Privacy-Preserving Identity Transactions online: The use case of Brokered Identity Federations

by

Maryam Shahid
B.S. Computer Sciences,
Lahore University of Management Sciences (2018)

Submitted to the
Institute for Data, Systems, and Society
and
Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degrees of
Master of Science in Technology and Policy
and
Master of Science in Electrical Engineering and Computer Science
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
February 2021

Author_____
Technology and Policy Program
and
Electrical Engineering and Computer Science
January 15, 2021

Certified by_____
David Clark
Senior Research Scientist, Advanced Network Architecture
Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Certified by_____
Karen Sollins
Senior Research Scientist, Advanced Network Architecture
Computer Science and Artificial Intelligence Laboratory
Thesis Supervisor

Accepted by_____
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Accepted by_____
Noelle E. Selin
Associate Professor, Institute for Data, Systems, and Society and
Department of Earth, Atmospheric and Planetary Sciences
Director, Technology and Policy Program

Privacy-Preserving Identity Transactions online:
The use case of Brokered Identity Federations

by

Maryam Shahid

Submitted to the
Institute for Data, Systems, and Society
and
Department of Electrical Engineering and Computer Science

on January 15, 2021, in partial fulfillment of the
requirements for the degrees of

Master of Science in Technology and Policy
and
Master of Science in Electrical Engineering and Computer Science

Abstract

Disinformation campaigns, created by fake accounts on online community platforms, have grown into one of the biggest threats against democracy, freedom and user perception of the truth. This thesis argues that mitigating this abuse and establishing trust online does not necessitate exposing identifying information about users on such platforms. Examples of identity mechanisms used by current platforms to curb this bad behaviour are included in the thesis to showcase the gaps in current schemes. There is a need to remodel the flow of identity transactions to cater to both anonymity and accountability considerations. To that extent, this thesis presents a use case of Brokered Identity Federations as a means to implement a one-time blind *proof-of-existence* that establishes a real-person is behind an account without revealing any excess identifiable information. The proposed architecture incorporates *proof-of-existence* by leveraging and re-purposing already existing enterprises, amongst whom identity interactions are divided, to maintain user anonymity while ensuring accountability in terms of limiting mass account creation. Lastly, the thesis discusses key considerations to take into account before the proposed architecture can be successfully realized.

Thesis Supervisor: David Clark
Title: Senior Research Scientist, Advanced Network Architecture
Computer Science and Artificial Intelligence Laboratory

Thesis Supervisor: Karen Sollins
Title: Senior Research Scientist, Advanced Network Architecture
Computer Science and Artificial Intelligence Laboratory

Acknowledgments

I am deeply indebted to my friends and mentors who have made my short time at MIT so memorable and rewarding.

First and foremost, I am grateful to my advisors *David Clark* and *Karen Sollins* for their guidance, discussions and conversations that have not only shaped my work but my approach to research and beyond. I will be forever thankful for all your encouragement and trust. I am also thankful to the larger *ANA* and *IPRI* group for their support and the many interesting conversations with colleagues and friends.

Navigating through the maze of MIT would have been impossible without the support of *Barb, Frank, Ed, Noelle* and the rest of the *TPP* and *IDSS* staff. Thank you for making me feel at home and letting me be part of the TPP family.

It has been an absolute pleasure to get to know my wonderful *TPP ('19,'20,'21 and '22) cohorts*. Thank you for being the absolute best people to be around. I will always remember TPP lounge discussions (and later Zoom sessions) on topics ranging from the serious to the absolutely ridiculous. I feel especially lucky to have met *Lydia*, my constant companion in the many many thesis work sessions we have had together.

To my favourite editor and all-around support system, my husband *Zain*, thank you for being here with me through this thesis and my MIT journey. Lastly, I want to thank my family, especially *my parents*, who have always encouraged me to follow my passions.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

THIS PAGE INTENTIONALLY LEFT BLANK

List of Figures

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

There have been many instances providing proof of Russian involvement in the 2016 US elections. The extent of Russian interference was unlike anything the world had ever seen before. According to the data accumulated by United States Senate Select Committee on Intelligence (SSCI), the propaganda reached at least "126 million people on Facebook, at least 20 million users on Instagram, 1.4 million users on Twitter, and uploaded over 1,000 videos to YouTube" [DiR+19]. This propagation of disinformation has been made possible by the ability of bad actors to create fake accounts on social media platforms with ease. If thousands of fake accounts are pushing forth content with an agenda, 'real' accounts[1] will not be able to differentiate them from legitimate content.

To understand the depth of disinformation, let's take an example from the 2016 elections itself. Russian involvement in the the 2016 elections via social bot accounts as well as human accounts have been detected across multiple platforms to influence public discourse - by creating distractions, diverting attention and driving polarization. On June 23, 2016, when Brexit was voted in, Russia started prompting Texas Secession initiatives with the hashtag #Texit with many fake Instagram accounts such as @rebeltexas, @_americafirst_ and @mericanfury promoting it. Content created by Facebook Page 'Heart of Texas' created and pushed this narrative regularly and in the period of 2016 to 2018 it had 4.8 million shares! [DiR+19] 'Heart of Texas' grew so large that its follower count was more than the official Texas Democrat and

[1]Here 'real' accounts means online identities controlled by a human for legitimate purposes.

Texas Republican Pages combined [Mic19]. The impact is not limited to promoting a divisive ideology by giving it the illusion of mass support, but it also resulted in actual pro-succession protests across the state by using a Facebook feature to create Events with designated time and location for people to gather.

Quick definitions: [Var+17], [Ste19]

- **Disinformation**: The deliberate creation and distribution of information that is false or deceptive in order to mislead an audience.
- **Misinformation**: Information that is false but not spread deliberately.
- **Propaganda**: Information that may or may not be true that is designed to engender support for political view or ideology.
- **Social bots**: Computer algorithms that assume an online identity and produce content and interact with users.

Simple reactive measures to ban accounts based on behavioural patterns is not enough to remove disinformation-spreading bot accounts. The sophistication of social bots and their patterns have developed enough to emulate human behaviour and the possibility of removing the account of a real user is possible - something most online social media platforms want to avoid. In the case of #Texit and many others, multi-fold tactics are used to create the illusion of real users and legitimate content. The social botnet and the bad actors behind it tend to leverage most of the features available on a platform, that means that instead of just posting or liking material, they tend to comment with positive or negative reactions; on Facebook, features like Pages, Events, Ads, Poking, Private Messaging are all utilised. Moreover, fake accounts try to create *brand identity* by creating the same or similar accounts across multiple platforms. For #Texit, there was heavy engagement on both Instagram and Facebook, and content was shared cross-platform. This lulls real users into trusting such fake accounts and believing in their false legitimacy. Using memes instead of sharing text is another technique to establish trust as memes build an emotional "in-group" connection that further makes these fake accounts seem "part of our crowd" and it's easier for real users to relate with the content. Examples of #Texit memes which target people with right-wing sentiment are shown in figure 1-1 and figure 1-2. Different (albeit fake) accounts continuously sharing and re-sharing content on armed insurgency, division and secession reinforces the perception that such opinions are commonplace and results in organic sharing and misinformation as well.

Figure 1-1: Examples of right-wing propaganda on 'Heart of Texas' Page

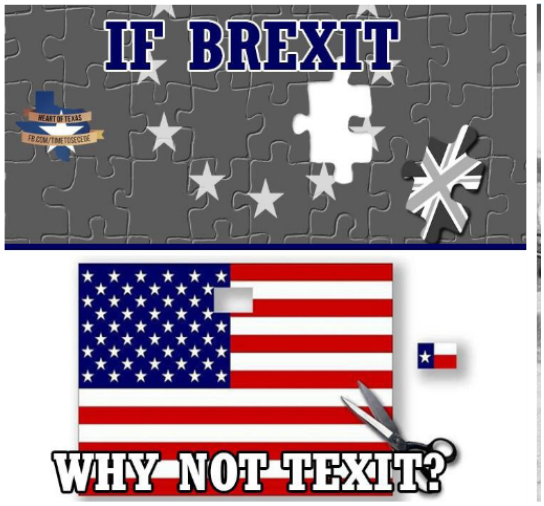

Figure 1-2: Example of #Texit propaganda

Organized abuse of digital identity online such as #Texit is not an isolated incident or limited to US soils. Russia uses social bots regularly to spread a pro-government agenda [How+19]. Brazil used similar techniques to drown out dissenting opinions during presidential campaigns in 2014 [Arn18]. China also used disinformation campaigns to undermine Hong Kong Protesters in 2019 [MM19]. Manipulating users by diversion, drowning out legitimate voices, amplifying organized propaganda messages and actively persuading users of polarizing content unwittingly erodes the trust placed in such platforms and the damage exceeds beyond the cyberworld.

Large social media platforms have been slow to counter the growing threat posed by this disinformation propaganda or sometimes even labeled as *Information Wars* aptly. Although efforts are being made by Facebook, Twitter, Instagram and YouTube, these efforts are more reactive than proactive in nature, and most times if not all, such actions impose a negative consequence on the privacy of all users involved or 'suspected' users at the least. The most recent move by Instagram is to introduce new authenticity measures which will "confirm who's behind an account when we (Instagram) see a pattern of potential inauthentic behavior. By prompting the people behind accounts to confirm their information, we (Instagram) will be able to better understand when accounts are attempting to mislead their followers, hold them accountable, and keep our community safe"[2]. The identification that Instagram can request includes either an image of birth certificate, driver's licence, passport or credit card among others. The purpose of the ID card will be to confirm the user's real name and age. While only name and age are required, digitally modified images to conceal SSN number, passport number or credit card are not acceptable. This is an urgent problem because Instagram, like many other social media platforms, is choosing the route of increasing accountability by means of identifiability and sacrificing anonymity/pseudo-anonymity in its stead. Anonymity and accountability are seen as opposite ends of a spectrum and an increase in one is equated with a decrease in the other. Moving towards a more accountable internet should not mean more sharing of users' personally identifiable information and increased tracking and surveillance.

[2]https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram

1.1 Ideal Model

The challenge in this space is to create an identity scheme, where both anonymity and accountability are ensured. Only real users are able to create accounts or bad actors are limited from creating too many accounts on an application, without compromising or exposing their actual identity.

To ensure accountability in terms of limitation on the number of accounts a user can create without exposing excessive personal attributes of the user, there is a need to divide identity tasks across domains and separate the role of identity verification from the application that a user wants to interact with.

In an ideal system, there should be two parties (at the least), let's call them server A and server B. Server A is the application that the user wants to create an account on, and server B will verify a user's identity. A user will first go to server B which will authenticate a user's identity by a stringently unique attribute, this could be their credit card number, their SSN number, or a combination of their real name, birth date and country. After server B verifies the user's identity, it will generate a token that the user will pass on to server A as proof of their real-world identity and be able to create an account. This token will have some good properties such as:

1. to be non-traceable, which means server A should not know the identity of server B and vice versa[3],
2. to not expose excess information, which means server A should know that a person, let's say Tooba, wants to join their website without knowing exactly what was used to prove Tooba is who he says he is at server B, and
3. to remember if the same user is making a request for the first time or not, which means server A should know Tooba is requesting for an account for the n^{th} time.

An example of the ideal model is shown in figure 1-3.

[3]It should be noted that this condition cannot be fulfilled in this 'Ideal' scenario since A is directly communicating to B and if A does not know B, there is no way to confirm that the token was sent from B. A close alternative would be that A 'forgets' that it received a token from B, after confirming the token according to necessary security standards.

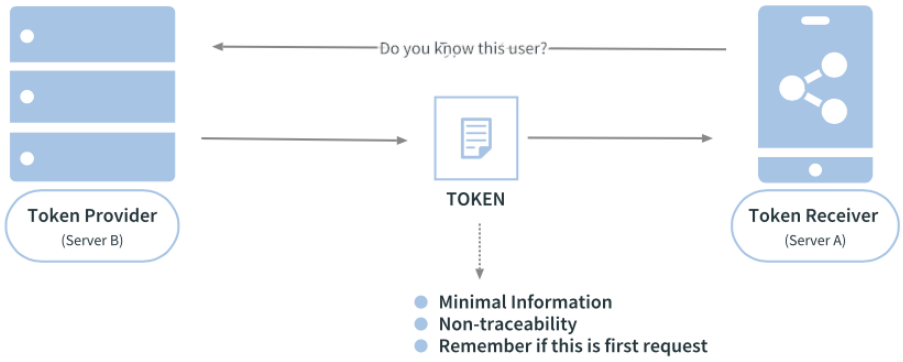

Figure 1-3: Ideal Model

1.1.1 Why is this hard?

To build the ideal system, there must be some motivation or business proposition for server A and server B (token receiver and token provider respectively) to act in a privacy-preserving manner. The ideal system requires both parties to 'forget' where user information is coming from and where the information is used, and we would have to take the word of both parties that they will behave in such a manner and not track the user. This is an impossible scenario since server A and server B as profit-maximising entities, would want more data of users to better understand their customers. Moreover, they might need to hold on to user data and metadata for audit purposes. On the other end, users and user advocacy groups would not be satisfied with having the 'word' of the parties involved in maintaining the trust that their information will not be used unnecessarily.

1.1.2 What can be done?

The issue of trust is fundamental in cyberspace. One workaround to establish this trust is to adopt Federated Identity Management which allows for a distribution of identity information between multiple domains who are in a partnership. There is clean separation between the service a client is accessing and the associated authentication and authorization procedure. One version of Federated Identity Management is the **brokered federation** in which a third

party acts as a mediator between the interactions of token provider and receiver. The main features of such a trust framework has 4 major components; the 'Identity Service Provider', responsible for identity proofing services (token provider); the 'Relying Party' which is the service being accessed (token receiver); the user who is trying to access a service; and an 'Exchange' or a broker that provides a double-blind capability where the Identity Service Provider is not aware of who the Relying party is and vice versa. Here, the business model of the Exchange is based on how well it can mediate information between the Relying Party and Identity Service Provider so there is no conflict between retaining large amounts of data v. keeping interactions privacy-preserving.

1.2 Thesis Organization

The aim of this thesis is to make a use case of Brokered Identity Federations for the explicit purpose of limiting account creation by bad actors who contribute to organized digital identity abuse online.

First, the thesis expounds on the problem of organized digital identity abuse and how Social bots are the primary modes of conveyance for massive disinformation campaigns (chapter 2). Chapter 2 also discusses existing solutions, in addition to social bot detection techniques, to limit fake account creation and their shortfalls. The thesis does not have a traditional "Related Works" chapter and relevant work is discussed in situations where it becomes relevant.

Chapter 3 addresses the anonymity-accountability debate. The false notion that higher accountability measures result in lower anonymity exists because accountability and anonymity are viewed as a spectrum rather than a quadrant where it is possible to preserve both. The thesis makes the case that instead of dividing value we can create more value if we change the expected architecture and identity transaction flows.

The thesis includes a proposed architecture based on previously established Brokered Identity Federation systems but re-purposed to limit fake account creation and tackle the rise in Social bots. It identifies pain points and further considerations to take into account that make the architecture work and expounds on security and privacy requirements that need to be fulfilled. This is covered in chapter 4, chapter 5 and chapter 6.

Finally, Chapter 7 lays out some key conclusions while Appendix B sketches out a working

example of privacy-preserving identity interactions expected through the proposed Brokered Identity Federation.

Chapter 2

What's the problem?

2.1 Organized digital identity abuse

To understand how we can limit access of bad actors to any community platform online, we need to first understand the characteristics of the bad actors and the harms they can cause. One such major characteristic is to classify bad actor accounts as bots v. non-bots. Bot accounts belonging to a single botnet/bot-family are controlled by one bad actor (bot-master). These bot accounts aim to push a narrative to influence the conversation on the community platform[Mur+16]. On the other hand, in the case of non-bot accounts, a bad actor returns with another digital identity after their first (or first couple) digital identity gets banned or blocked from an online community platform. This is troublesome because these bad actors usually target individuals and create victims of cyber harassment. In 2019, the app TikTok made the news for exposing minors to predators when a man made multiple accounts to gain personal information about under-age girls to get their home addresses [TCH19]. Since the cost of making multiple accounts is so low, it makes the audience more vulnerable to the malicious behavior of bad actors.

2.1.1 The case of Social Bots

In this thesis, I will focus on reducing the impact of bad actors behind bot accounts rather than the non-bot harasser case. Online community platforms such as Facebook, Twitter, and

Instagram are now an essential part of communication and even news dissemination. Bad actors, individuals or organizations, have been leveraging these platforms to gain influence by use of these bot accounts. Bot accounts that are used online to imitate human behaviour are called Social Bots[Var+17] - defined as "computer algorithms that produce content and interact with users". Other terms such as Social Media Bots or Sybil bots are sometimes used to describe the same or similar phenomenon[Ora+20]. By their definition, Social Bots are not only used for malicious purposes. Benign Social bots exist that are used for posting tweets of earthquake alerts[Aki11], chatbots for customer service and news bots[Hei] as well.

However, there has been an increasing trend in using Social bots to subvert conversations online, deceive users and reduce the overall integrity of information available on these platforms. Approximately 8.5 % of Twitter users are bots as disclosed by Twitter[Sub+16]. Another study on Social bots shows that out of all English-speaking active users on Twitter, 9% to 15% exhibit bot-like behaviors[Var+17]. No one truly knows how big this number might actually be. According to an Oxford study, there is a strong evidence that organized social media manipulation campaigns have taken place in 70 countries in 2019[BH]. All this data goes to show that Social Bot accounts are one of the most important, if not the most important, cause of security threats to community platforms online.

Social bots have far and wide impacts. They have been used to infiltrate political discourse, steal personal information, spread misinformation and even affect the stock market. Social bots have been known to interfere in US midterm elections in 2010 and then again in US presidential elections in 2016. According to a study, one fifth of the conversation was produced by bots during the 2016 election[BF16]. Furthermore, bots have interfered in conversations about vaccinations with the purpose of creating highly polarized views[Fer+16], they have orchestrated successful campaigns about a tech company Cynk, which resulted in a 200-fold increase in the company's market value with valuation at $5B. By the time the bot activity was discovered and trading halted, the losses were already made[Fer+16]. Table 2.1 shows a comprehensive list of types of Social bots and their purpose.

Cashtag piggybacking bots are the most recent type of social bots discovered and there might be more unknown attacks already in play. If these attacks continually succeed and users are unable to differentiate between a fake user and a real one, it will significantly impact how these platforms are used. Instead of bastions of free global speech, they will become the strongest media to disseminate disinformation. User trust will deteriorate, the world will

Social Bot Type	Description
Cashtag piggybacking bots	Social botnet that promotes low-value stocks by exploiting the popularity of high-value stocks [Cre+19]
Astroturfing bot	Bots that create the appearance of widespread support of an opinion or a person [Rat+11]
Social botnets in political conflict	Bots aimed at deflecting readers, drowning out dissenting opinions in political environment. [AYM15]
Influence bots	Highly sophisticated bots that emulate human behaviour and hence can shape discussions [Sub+16]
Infiltration of an organisation	Bots that aim at becoming friends with employees of an organization to gain potentially harmful information [Ely+16]
Sybils	User accounts use for a disproportionately large influence [AlR+15]
Doppelganger bots	Clone actual users and mimic human behaviour [GVG15]
Spam bots	Bots that spread malicious links, send unsolicited messages, crowd conversations and hijack trending topics [Wan10]
Fake accounts used for botnet command and control	Accounts that share encrypted commands for a botnet attack. [SAV14]

Table 2.1: Malicious Social Media Bots

become more polarized and bad actors will easily ruin or build reputations by altering the perception online.

2.1.2 Existing Social Bot Detection Techniques

Since the impact of malicious Social Bots can dramatically impact the discourse online, there has been extensive work done in detecting Social Bots and differentiating them from real users. One taxonomy proposed in literature[Fer+16] divides the detection techniques into four main classes:

- bot detection based on network information,
- system based on crowd-sourcing,
- machine learning method based on differentiating features, and
- hybrid methods

Network information based detection systems assume that Social bots within a bot-family will interact with each other for the majority of the time and have much smaller number of links with real users. Facebook Immune System[SCM] and SybilRank System[Cao+12] exploit this feature to detect entire bot-families. However, this can be easily circumvented by having

links similar to real users and mimicking these interactions. Wang et al.[Wan+12] proposed using real humans to check whether an account and its content depict a real user or not, and then crowd-source the results. The strategy showed a near zero false positive rate so it was successful, however the drawback is the usage of too much manpower which makes this method unsustainable. Using machine learning to detect suspicious behaviour patterns is a popular technique to detect Social Bots - one example is the open source BotOrNot system[Dav+16] which was a first of its kind technique that used supervised learning to score a detection accuracy rate of 95% on the 2014 dataset it was tested on. However, with changing Social Bot behaviour, the same ML model might not be sufficient for sole detection purposes. Hybrid methods try to combine the previous approaches to get a more holistic picture but results are usually varied.

Despite the plethora of bot detection techniques, the problem is still ever present. One of the reasons is that this research area is relatively new, the effects of malicious Social Bots were recognized only in the last 10 years. Secondly, the data-sets required to run experiments are available to limited researchers and hence reduce the solution-space. Moreover, detection of Social bots online is becoming more and more difficult as the sophistication of the bots themselves is increasing, making it hard to distinguish between a bot and a real user. Social bots now fill their profiles from information from the web, post content at irregular times and can even hold a conversation on multitudes of topics. Some bots even "clone" the behaviour of legitimate users to evade detection.

There have been many studies that have run evaluation on these detection techniques and have reported critical findings about holes in detection methodology and inefficiencies in current techniques. An experiment[Ely+13] targeting employees of certain organizations (mostly tech firms who realize the consequences of information leakages online) were able to infiltrate users with a success rate of approximately 50-70%. Here, infiltration is defined as accepting a Social bot's friend request. In another experiment[GAA18], scientists were able to show successful evasion of the popular bot detection scheme BotOrNot, when it did not detect a fairly sophisticated botnet that was trying to push a certain propaganda. It seems that bot detection and sophisticated bot creation will remain a cat and mouse game where bot detection techniques are improving but so are bot evasion techniques.

Detection and removal of fake accounts is a very reactive approach to this burgeoning problem. There is a need to take pre-emptive action to limit the creation of fake accounts. As of right

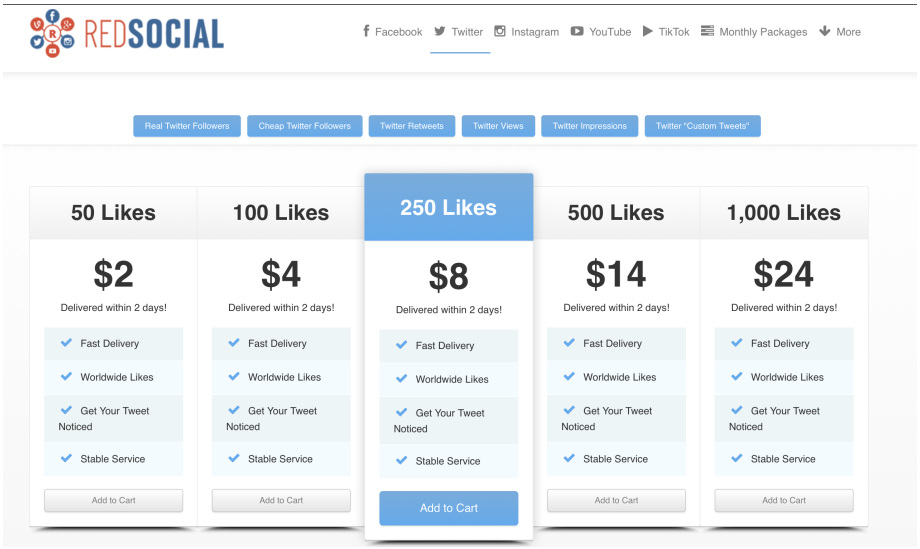

Figure 2-1: Example of a web service to buy fake likes

now, buying accounts for fake follows and likes is a business that botmasters are running smoothly. One does not even need to go to the Dark Web to avail these services; a simple Google Search promises deals such as '50 twitter likes for $1'. An example website is shown in figure 2-1 [1]. The associated cost of creating a new account on many of these online community platforms require only an email address to sign up.

In my thesis, I propose a scheme to limit the abilities of bot masters to be able to create thousands of fake accounts without compromising on user's privacy constraints. It is also important to discuss current solutions used by platforms to reduce fake account creation and their limitations.

2.2 Existing solutions to limit fake account creation

2.2.1 Reputation Schemes

One way to reduce fake account creation is to reduce the value attached to a new account, or conversely increase the value of an old account which has received ratings or votes. Con-

[1] https://www.redsocial.net/buy-twitter-likes/

solidated ratings and votes are defined as reputation of an account[Vos04]. This reputation is tightly coupled with an identity - this is the identity of the account itself. And it is a non-tradable asset due to this tight coupling. In this section we will discuss two examples of existing reputation systems, their designs, and their flaws. While these examples are not exhaustive in terms of reputation schemes available, they provide a good foundation of how reputation schemes in the context of online community platforms behave.

Ebay

Ebay is one of the earliest reputation systems. The need for Ebay arose when users needed to purchase items from strangers online and were concerned about the product or service that would be delivered. This phenomenon was completely different from traditional auctions as you neither know who the sellers are, what their characteristics are, what others have claimed about them nor their location beyond their city. Hence, this would create little or no accountability among sellers and the incentive to cheat the buyers would increase.

Hence, Ebay introduced a system to score sellers based on their past interactions, i.e., their past sales. After each transaction, the buyer and seller can rate each other and give a positive, negative or neutral rating and provide brief feedback in the comments section. Furthermore, previously newcomers on the site were differentiated from experienced users by having a small sunglasses icon next to their username for the first month of membership. This ensured that users that had racked up a lot of negative reputation could not simply shed their old identity and come back to the platform without building a new reputation from scratch[Res+06].

However, there was little else done to monitor abuse of the reputation system itself. The motivation to abuse the system is that sellers receive higher financial gain from having a better reputation. According to an empirical study, established sellers could charge an 8.1 percent higher price than new sellers offering the same merchandise.

There have been cases of abuse when users sell or rent their high reputation identity to others to cheat buyers and pretend they have a much higher reputation than they actually do. Sellers have also improved their reputations through notional transactions whose purpose is to raise their positive feedback. The cost of abusing such a reputation is relatively small, whereas the advantage can be quantified in monetary terms[BM06]. Fake accounts can be created manually but using botnets allows bad actors to scale their efforts. All it takes is assembling an

attack script that inputs data into a registration form and then using bot networks to distribute the script and create many fake accounts.

In other reputation systems, the case can be even worse. For example, for Yelp, you don't actually need to make any purchase to review a restaurant and it is hard to tell whether a review is legitimate or not. "Review gaming" is a technique in which bad actors extort money from small business owners, threatening to tarnish their Yelp reputation by bombarding them with fake bad reviews[Kan07]. Many businesses have no option but to oblige since a small drop in ranking can result in significant business loss to the owner. The problem here is that one person has the influence of multiple 'fake users' and hence can manipulate the discourse online. Yelp has its own proprietary filtering algorithm to deal with suspicious or fake reviews because the problem is so prevalent; nearly 1 out of 5 reviews is marked as fake by Yelp[LZ16]. Such an aggressive method might purge fake reviews but since there are no checks on Yelp's own algorithm, sometimes legitimate reviews also get purged. While this is a risk Yelp is willing to take, it might not be the case for other online community platforms.

StackOverflow

Other than auction sites such as Ebay, there are also Q&A sites such as StackOverflow, where reputation schemes are used to assign value to an account. Questions on StackOverflow are usually more technical in nature and revolve around coding, math questions and other technical topics and hence it is quite popular among students, educators and industry experts.

Good reputation is earned by asking good questions, commenting, upvoting or downvoting answers based on their quality and most importantly answering questions. The platform provides motivation for expert users to provide answers to more questions (high reputation equals high rewards and better access for them), and limits behaviour of bad actors (low reputation limits ability to post questions, answer questions and even comment sometimes).

Unlike Ebay where reputation is an add-on to increase credibility of sellers, and hence make the platform more attractive, in StackOverflow reputation is the key underpinning which sorts information into "more helpful" or "less helpful" category. Instead of reputation being used to mark trustworthiness of a user, it marks the trustworthiness of the content - the answer provided by a user. A user on StackOverflow doesn't care whether the question was posted by a credible user or answered by a credible user, it only cares about the upvotes or downvotes

an answer has and whether it actually solves the question asked. However, reputation does determine the ability of a user to interact with the platform. The higher a user's reputation, the more privileges they have, such as getting moderator status and being able to 'close' a question as low quality.

The abuse scenarios possible on StackOverflow by a social botnet are either targeting an individual user and serially downvoting them for them to lose reputation unfairly or flagging certain content (questions or answers) as unhelpful, duplicates or inappropriate.

This loss of reputation could be harsh to long-time users but outside this very specific community, it does not have a lot of real-world damage. Being able to manipulate content on a Q&A that is hyper specific to only a given area of expertise, that does not hold any user Personally Identifiable Information (PII) and provides little gain for the abuser means that there is not a lot of focus put into verifying whether an account (or a user) is real or not. However, spammers still exist and StackOverflow has various built-in mechanisms that deal with such behaviour. The Penalty Box is one of such mechanisms in which a user's reputation drops to 1 (the lowest possible reputation), their profile is marked as suspended and they lose their ability to vote, comment, ask a question or answer a question [Atw09]. This Penalty Box is based on a user's anomalous behaviour rather than a group of colluding users.

2.2.2 'Real' attribute schemes

Like Reputation Schemes, real attribute schemes are another way to limit fake account creation. They differ in how they try to limit fake accounts. Instead of attaching value to an account, they associate an account with a user's Personally Identifiable Information (PII). Since PII are roughly unique to an individual they are harder to fake. Examples of PII are real names, citizenship information, phone numbers, etc.

The biggest issue with using PII is that they pose a huge privacy concern, allowing service providers to gather identifiable information about users and compromising their anonymity not only on their platform but to all the data partners/marketers they share user information with.

Twitter and Phone Numbers

Twitter, like most other leading online community platforms, used only an email address to sign up new users. However, creating fake email accounts takes little effort and cost. Automated bots can easily be used to exploit applications that verify email accounts, hence creating an unlimited number of phony accounts on a platform. In recent years Twitter has faced a lot of backlash on abuse, fake followers and fake likes on their platform.

To counter this negative publicity and lack of user accountability on the platform, Twitter has shifted to asking for emails and phone numbers. This allows them to verify phone number ownership during sign up by sending verifying codes via text messages to new users. Phone numbers are much harder to fake, especially in large numbers (unlike email), since there is a need to physically purchase a SIM and depending on where the bad actor is, purchasing SIMs can provide moderate to high difficulty - in some countries, SIMs are connected to government issued identity of a user. Moreover, phone carriers can provide information to further validate the legitimacy of a number and will show if, for example, the number is out of the country of origin (is roaming abroad) or has moved to another phone. Connecting an account with a single number ensures that it cannot be used for multiple accounts and especially for accounts that have been banned before.

However, using phone numbers to differentiate between a legitimate user and a fake user might not be an effective solution because of paid or even unpaid services such as Pinger, TextNow and smsrecievefree.com (figure 2-2) which provides you many permanent fake numbers or access to numbers to receive a text messages to verify account creation. This makes the account to phone number mapping virtually useless. In the case of Twitter, this technique cannot bypass account verification because Twitter blocks numbers that are VOIP-based, which is what these services provide.

The biggest issue with using phone numbers to verify accounts still remains the lack of privacy and possible misuse of phone numbers to track users. In 2019, Twitter issued a public statement admitting that they inadvertently allowed marketers to target ads based on phone numbers provided for account security[Sul19]. It also allowed marketers to build a better advertising profile by using such PII. Since, PII are fixed for people even outside the twitter-sphere, this means that companies can track users with their phone numbers through a myriad of public records such as voter registration, real estate transactions and marriage records - and correlate

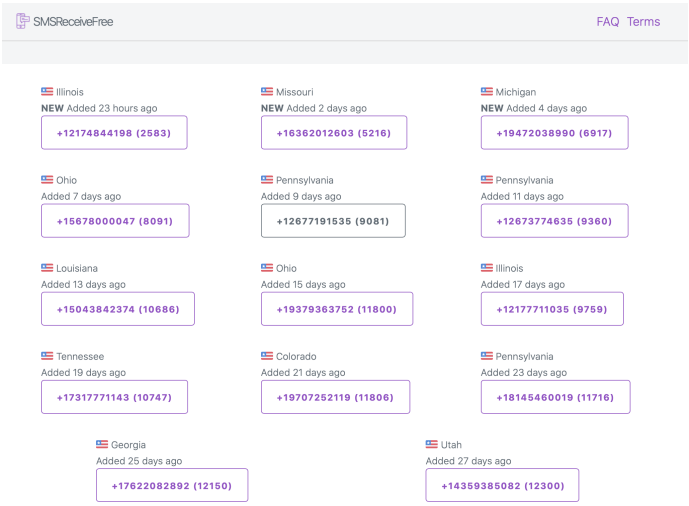

Figure 2-2: Example of a web service to bypass phone verification

that information with their Twitter data.

Facebook and Real Names

Facebook's "Real Name" policy requires users to use their legal name when creating their profile on the platform. This policy was always there as part of the terms and services that each user agreed to while signing up, however it started being enforced sometime in 2014 with mass account deletions if the name used in profile was not their real name. According to Facebook, the policy is there to protect users as there exists proof of online harassment being largely anonymous.

However, there has been speculation that the biggest reason for the push to de-anonymize people on Facebook is the financial stake Facebook has in user identity. 69% of Facebook's revenue comes from marketing targeted ads[Wil15]. Fake profiles mean lower quality data and hence affects their bottom line. Moreover, when Facebook reported that around 8.7% of accounts on their platform were fake in 2012[Tho12], it led to one of the largest drops in share prices. To combat this lack of faith in investors, Facebook started cracking down on fake profiles. Hence, providing PII at Facebook was more related to the tracking users rather than maintaining accountability of bad actors.

There was a strong backlash to Facebook's policy as it undermined the LGBTQ+ community (drag queens were not allowed to keep their drag names), Native Americans (their names unjustly not considered as their real names) and even abuse victims (who used pseudo-names to keep away from abusers)[Vaa15]. Activists claimed that the policy is superficial in identifying fake accounts and there are no reports (from Facebook or otherwise) that show that providing legal names curbs bad actors online - it instead puts marginalized communities more at risk[War15]. In response, Facebook did soften the policy but did not back down and according to Electronic Frontier Foundation (EFF) the changes are still not enough to protect online communities[HB15].

Second Life and Credit Card Information

The case of Second Life is unique when it comes to online community platforms. Where Facebook and Twitter are seen as an extension of your social circle, Second Life is a virtual world where the large majority of the users want to disconnect from their actual identity. Users create avatars with pseudonymous names, gender, profession, age and any other attribute they can change in their online profile. As the name suggests, users live a second, separate life from their real-world identity. Hence, anonymity and privacy are fundamental aspects of the virtual world that ensure users can have virtual weddings, can work as virtual sex workers and live a completely different virtual life without the danger of their real-life identity interfering or exposing them.

Second Life is noteworthy in a sense that it did not have a prolific problem of fake accounts or fake users. Neither did it face any backlash on accountability mechanisms being too lax (the case of twitter introducing phone number association to accounts to reduce bot accounts) or unnecessarily stringent (the case of Facebook softening policy on real names scheme). This subsection is important because it highlights how a profit-making business removed accountability measures to reduce the burden of entry for new users and the subsequent online community's reaction to this change.

When Second Life was launched, a user could sign up as a free account holder or a premium account holder. Premium account holders had the capability to buy land using Linden Dollars which could be exchanged for US dollars. Hence, each premium account holder provided their credit card information for such a transaction. However, each free account holder had to provide credit card information as well, not for any monetary purposes, but to use a valid

credit card number as a form of "identity verification". Fundamentally, this meant Linden Lab could track every user to their real identity through their credit card information and while the users were pseudo-anonymous to each other, such was not the case for Linden Lab itself.

However, in 2006, Linden Lab decided to forgo this requirement for free account holders as they realized half of the people who initiated the registration process stopped at the stage where their credit card information was requested. This made sense as some people might not be willing to provide this private information, while others might not have banking credentials to sign up with. Changing to just username and password for free account holders would reduce the barrier of entry and it did as the new simple sign up process led to a fourfold increase in the rate of new users signing up for Second Life[Boe08]. This highlights an important aspect of why many online social platforms refuse to add more accountability metrics when users sign up, i.e. to encourage a higher sign-up rate, even if it means embracing the potential dangers of social botnets' creations.

Many long-time users or 'residents' protested when this change was announced. These residents joined together and created Proposition 1503 which lobbied Linden Lab to revert these changes and have same identity verification requirements as old users[Boe08]. For a community that heavily put emphasis on anonymity, this seemed like a strange response. However, the online community at large felt that without the accountability measure of providing credit card information, misbehaviour in the online world would increase - a phenomenon labelled as "griefing". Anyone could create multiple accounts and feel no repercussions of being banned after bad behaviour. So, it was this user-based governance that pushed for more accountability measures. Linden Lab in response announced the following changes: "Each resident's profile now includes a field revealing ... one of three status entries: (1) 'No Payment Info on File'-account was created with no credit card or Paypal; (2) 'Payment Info on File'-account has provided a credit card or Paypal; (3) 'Payment Info Used'-credit card or Paypal on account has successfully been billed. We plan to provide features in future updates to mark specific parts of the Second Life world (or allow residents to mark their own land) as accessible only to accounts with payment information"[Boe08].

The creators of Second Life, Linden Labs, launched this virtual world in 2003, and despite it no longer being a cultural phenomenon it was in the 2000s, Second Life has still around half a million active monthly users[Bus20] and during the pandemic, the virtual world has seen a 60% increase in new users[Kar20]. Hence, the mechanisms of control observed in Second Life

are still relevant for online community platforms today.

2.3 Why these current schemes don't work

As discussed previously, Social Bot Detection technique leaves a lot to be desired. The effectiveness of the solution is incomplete - as Social Bots become better at evading detection techniques, the solutions become obsolete. Other methodologies to reduce the impact of fake accounts or bad actors online have also not been able to rid the internet of disinformation and its influence campaigns. Most solutions are platform-specific, which means that no one other than the platform itself can comment on the efficacy of such techniques. Online community platforms such as Yelp and StackOverflow are willing to remove content which they deem suspicious even if it might be a harmless legitimate user and are satisfied with their overzealous approach. On the other end, techniques used by larger platforms have not been sufficient to counter the threat of fake account creation. Facebook, Instagram and Twitter are three of the largest soocial media platforms used for dissemination of disinformation despite the many ploys they have applied to curtail such threats. One could be optimistic and say that they do not want to face the backlash from users who are unjustly blocked by algorithms who think their behaviour is suspicious, or one could be more cynical and claim the slow response is due to the fact that curbing disinformation hurts their bottom line. A study carried out by MIT researchers found that on Twitter, false news being re-tweeted (re-shared) is 70% more likely that true stories, after all a sensationalized article is much more interesting to read than one which does not have as much entertainment value [VRA18]. Since, these large online social media companies rely on user engagement to generate revenue, it might seem counter-intuitive to reduce this engagement. In scenarios of conflicting motivations such as these, it is hard to distinguish between actual security practices and security theater[2]. There is a need to distribute the responsibility of verifying the legitimacy of a user from the service which the user wants to access. But first there is a need to come to an agreement on what online identity, anonymity and accountability entail and why it has been hard to provide identity mechanisms that preserve both anonymity and accountability.

[2]Security theater: practice of taking security measures that are intended to provide the feeling of improved security while doing little or nothing to achieve it.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 3

Anonymity and Accountability

A common misconception among users of online community platforms, policy makers and even technologists is that one can have stronger anonymity *or* better accountability - but not both. One major reason for the perpetuation of the fallacy of "anonymity *or* accountability" is the lack of agreement on the definition of accountability, which impacts accountability research and how its is perceived. The role of identity is central in this dilemma. Identity has many definitions, varying on the socio-political contexts, however what is important to understand is that an individual's "identity" and "online identity" may have very little overlap. The Internet Society explains online identity as "...the sum of your characteristics and interactions. Because you interact differently with each website you visit, each of those websites will have a different picture of who you are and what you do"[1]. Later in the thesis (Chapter 5), the definition of online identity will be further restricted to limit to the context of the proposed architecture.

Hence, it is first important to understand how anonymity and accountability are defined, especially in the context of applications and services online. Only then can the false dilemma of *anonymity v. accountability* can be discussed and the possibility of *accountability with anonymity* be introduced.

[1]https://www.internetsociety.org/wp-content/uploads/2017/11/Understanding-your-Online-Identity-An-Overview-of-Identity.pdf

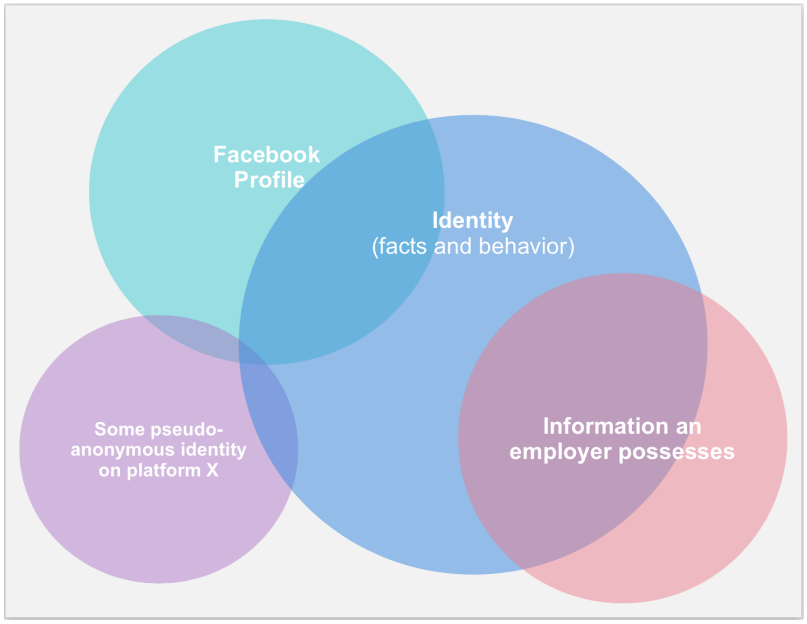

Figure 3-1: Overlap of identity and partial-identities online

3.1 Defining Anonymity

Anonymity is defined as being unidentifiable within a group called the *anonymity set* [PK01]. The strength of anonymity grows as the size of this anonymity set increases as well as the similarity between the attributes of the members of the anonymity set. In the context of the internet, anonymity acts as a privacy enabler, allowing the user's natural identity to be unlinked to certain degree from their online persona. The level of unlinkability depends on the design and protections offered by the online applications that an individual uses. Figure 3-1 shows a possible overlap of online and offline identities and depending on what services are being used, the level of personal information varies.

According to sociologist Gary Marx, to be fully anonymous an individual must be unidentifiable across 7 identity dimensions, which are legal name, location, pseudonyms related to name or location, pseudonyms not related to name or location, behaviour pattern knowledge, social categorization and skills that reveal personal characteristics [Mar99]. The thesis does not take a stance on whether a user should be *fully anonymous*, instead it leaves that to the context

of the platform involved. Some applications like 4chan, provide anonymity across all these 7 dimensions, others like Twitter provide a relative amount of pseudo-anonymity where users enjoy unidentifiability across 5 or 6 of the aforementioned dimensions, and the user can choose to reveal more information about themselves if they want[2].

3.2 Defining Accountability

There exists myriad literature on what the term *accountability* entails and many existing formal frameworks for accountability[Fei+11]. In the offline world, the core meaning of accountability is "the process of being called 'to account' to some authority for one's actions"[Mul00]. However, according to Mulgan the term has expanded to include new aspects such as "responsibility", "internal control", "responsiveness," and "dialogue"[Mul00]. These expansions are significant because they allow nuances that are applicable to the online world. On the internet, the policy enforcers are internal rather than an external judicial body, sometimes it is difficult to hold every user accountable for minor crimes and a more passive rather than active approach is possible. Hence, in the technical context, the definition of accountability was championed by Lampson as: " (Accountability is) the ability to hold an entity, such as a person or organization, responsible for its actions"[Lam05]. Although useful, Lampson's focus on the "entity being held responsible" relies on the identification of the entity. It is Feigenbaum et al. explicit focus on "be[ing] punished" that allows for accountability being achieved without the level of identifiability that is typically assumed to be required [FJW11]. This decoupling of accountability from identity makes Feigenbaum's notion of the possibility of "sanctions, holding responsible, or punishment" the most relevant to this thesis. Especially automated punishments, where potential violations of privacy are deterred by the prospect of negative consequences [Fei+11].

3.3 Anonymity-Accountability Axes

Not having an agreed upon definition of what accountability and online identity entail is the major reason behind platforms being unable to incorporate both accountability and anonymity in their systems. The misconception that anonymity and accountability are a zero-sum game

[2]the level of unidentifiability has much reduced since Twitter started requesting phone numbers of users who have 'supposedly' violated the Twitter terms and agreements

and hence both cannot co-exist is the reason that the sliding scale framework exists in prior literature. This sliding scale is depicted in figure 3-2[Wol12].

On the other hand, when we accept that accountability does not always require identifiability, the one-dimensional model becomes unnecessarily restricting. Wolff reports an alternative two-dimensional framework which allows for different combinations of accountability and anonymity to exist[Wol12]. Figure 3-3 shows this alternative 2D viewpoint populated with platforms that have been discussed in chapter 2 among others[3]. Visualising the anonymity-accountability axis as two-dimensional emphasises the possibility to keep both anonymity and accountability at high-levels. This possibility is not just hypothetical as the figure shows actual platforms that have managed such combinations of accountability and anonymity. These platforms punish users for their bad behaviour but do not expect additional PII from the user, other than what is necessary for the functionality of the service.

The extreme top left depicts platforms that require high accountability and high identifiability as well. These platforms include healthcare, government services and banking - applications which have a much higher security risk if hijacked by malicious actors. Moreover, services such as these *need* to know highly personal and sensitive information about an individual to fulfill their functionality.

On the opposite end, the extreme bottom right depicts platforms that have much lower accountability and higher anonymity; applications such as 4-chan and 8-chan fall in this category where there is no user-profile creation and each interaction with the platforms is independent of each other. The low accountability measures has allowed such platforms to be abused and has been complicit in hate crime, nude photo leaks and hacker groups, even though the aim was to build a message boards where people of similar interests could share their thoughts[4].

Eight years ago, the threat posed by social bots as well as their sophistication was not advanced enough to cause major concern. Hence, Wolff categorized online community platforms such as Facebook higher in both the anonymity and accountability axis than the current figure 3-3. The lack of a cohesive and effective response to punish bad actors on Facebook, Twitter and Instagram has lead to their placement in the bottom left quadrant. These major social media

[3]While the dimensions of the 2D structure are lifted from Wolff's thesis, the placement of the platforms themselves is altered based on the current information collected for this thesis.

[4]https://www.cnet.com/news/8chan-8kun-4chan-endchan-what-you-need-to-know-internet-forums/

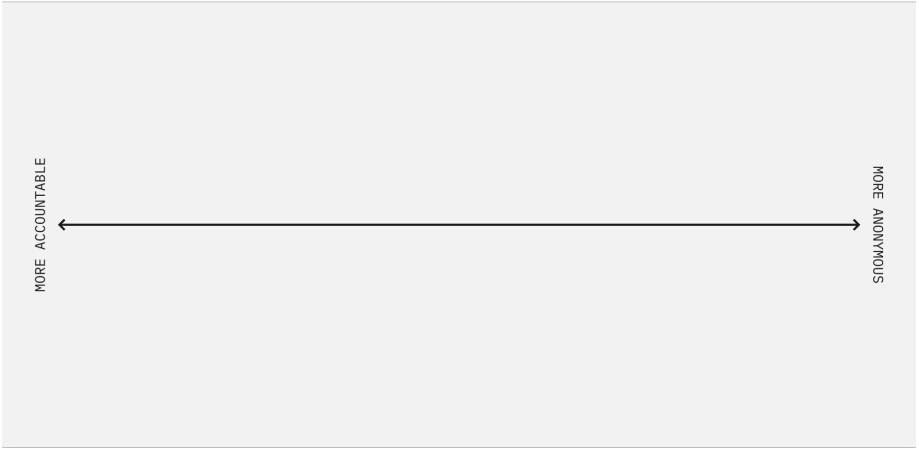

Figure 3-2: Anonymity and accountability incorrectly viewed as a zero-sum game in one-dimensional framework

platforms have low accountability measures as they have been unable to deter bad behaviour created by social bots and other bad actors. Yet they continuously expect higher and higher identifiable information from users, such as their phone number or in some instances their ID information which results in a weak anonymity.

Big platforms such as these are the ones most in need for the architecture proposed in this framework which will be discussed later in chapter 5.

The two-dimensional axis is not intended to rate one platform's accountability-anonymity measures as superior or makes any claim of adopting a platform's methodology as the best. Instead it describes how it is possible to have secure privacy-preserving interactions online where the applications necessitate different degrees of anonymity and accountability.

3.4 Accountability by limiting account creation

Previous chapters discussed the considerable consequences of allowing fake users to abuse online community platforms, which includes, but is not limited to, creating fake reputation, drowning out dissenting opinions, impacting political discourse and even exploiting the stock exchange to create fraudulent market value. Chapter 2 also examined the current measures to detect and limit bad behaviour online by holding users accountable for their actions, and the

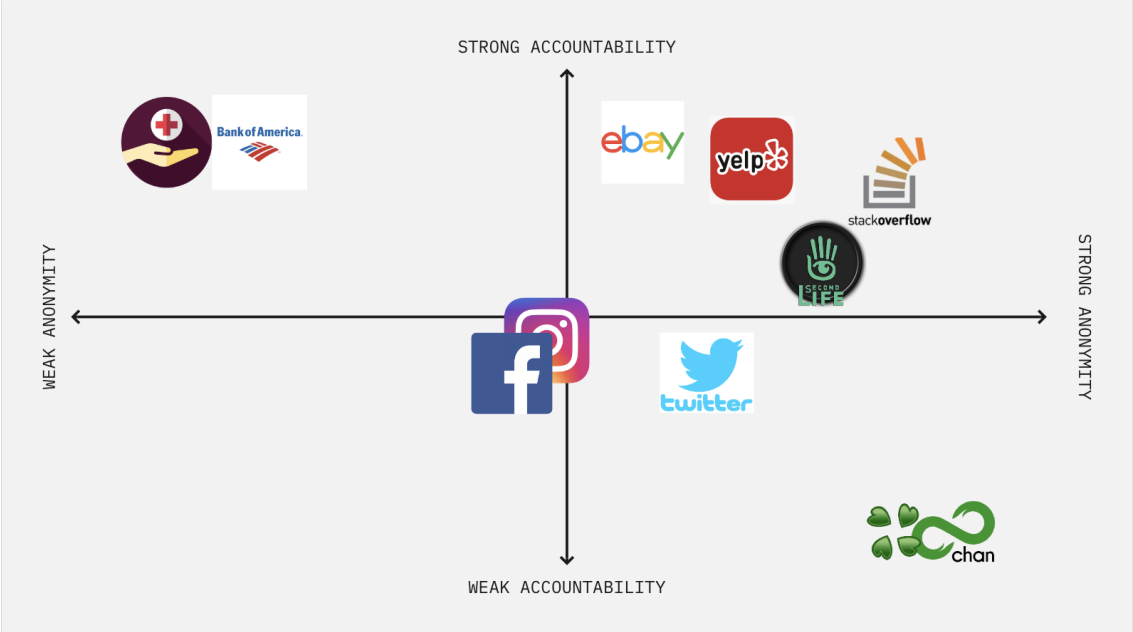

Figure 3-3: Anonymity and accountability in alternate two-dimensional framework

effectiveness of such solutions. Social Bot detection techniques and Reputation schemes rely on user behaviour patterns to categorize users as "malicious/fake" or "non-malicious" - and block their content or/and account based on the severity of their transgressions. However, due to the (pseudo-)anonymous nature of most online social media platforms, anyone can create more accounts to make up for their blocked accounts - the only downside being the loss of built up reputation and following. On the other hand, "real" attribute schemes rely on unique personally identifiable information (PII) about a user for attribution purposes and block current and future accounts based on PII such as email, phone number, credit card information, etc. Releasing PII to online community platforms has a track record of being misused for non-accountability purposes with Facebook and Twitter leaking users' phone numbers to advertising companies [LZ16]. Moreover, PII can be used as a very strong identifier for an individual, for example, a phone number can be used to track a person's home address, past addresses in the last decade, full names of family members and even a criminal record just by accessing public records [Che19].

The problem of proper accountability measures is rooted in online identity transactions - what they are, how they are defined and how they are perceived by users, the platforms and policy makers. The discardable nature of online identity is one of the major reasons why punishments such as content deletion and account blocking is not as effective as an individual can just create another account. Worse still is that since the cost of new identity on a platform is so low (usually only needs an email address) - social botnets are able to create thousands of accounts by running a simple script.

Hence in this thesis, the path of limiting account creation is chosen by adding cost for creating a new account by providing proof-of-existence (chapter 4) and assigning punishment for creating too many accounts by blocking new account creation after a certain number of accounts threshold (chapter 5). By making online identities less discardable, the top right corner in the 2D framework can truly reach its full potential of having high anonymity-accountability combination.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 4

Blind Proof-of-Existence

In the previous chapters, it has been established that harms caused by organized digital identity abuse are widespread and damaging to the discourse on online platforms, extending to real-world consequences as well. The current solutions provide only a partial answer and are platform-dependent. The downside of platform-dependent solutions is that they are not applicable to other platforms and there is no accountability on how the solution works unless the platform releases the solution as open-source code. Often times, providing a privacy-conscious, noninvasive security measure is at odds to the profit-maximizing motivations of a data-driven platform. Hence, trusting these online social communities to be motivated to take strict actions and regulate themselves is not effective.

The question arises then, *how can we improve accountability without harming the privacy or anonymity of the users of the platform*? There is a need to introduce an entity that takes an identity-provider role and establishes trust with the platform (identity-receiver). The follow-up question then becomes *who can take the role of the identity-provider*? *And how do we establish such trust*?

In Section 1.1 I ran through an example of what an ideal model which preserves both anonymity and accountability would look like and why it is difficult to establish such a relationship. The identity-receiver, or the platform on which the user wants to create an account is called the Relying Party. Further details of what the Relying Party pertains to are provided in section 5.2.1. To maintain anonymity at the Relying Party and prevent data tracking using highly personal

information, it is vital that the Relying Party not be in charge of identity proofing. After all, there are no guarantees that the Relying Party will not use the information a user provides for more than identity verification purposes. Moreover, if we provide identity information to the Relying Party, they can link multiple accounts that a user has created - something a user might not want to happen. So, the Relying Party must outsource the service that provides proof of the existence of the user - this *proof-of-existence* is an assertion about the offline/real identity of the user. Distribution of identity information between multiple domains who are in a partnership is called an Identity Federation. By using an external service, the Relying Party cannot compromise a user's anonymity but will still prevent a social botmaster from creating multiple fake accounts. If the Relying Party is provided proof from a trusted party that a user is real, or a user is above 18, or any other requirement that the platform has, they don't really need to know the user's actual legal name or the user's actual date of birth - the proof from a trusted party should be sufficient.

This trusted party that provides *proof-of-existence* and identity proofing services is called the Identity Service Provider (formal definition and details are provided in section 5.2.1). Offloading identity verification services to an Identity Service Providers is not a rare phenomenon. Anytime a person uses their gmail identity to create an account on another platform, let's say to subscribe to The Washington Post, they use an Identity Service Provider. In this scenario, the Relying Party is The Washington Post, and the Identity Service Provider is Google. Not only does Google provide identity proofing services but also authentication services which means that every time a user signs into The Washington Post, they will use their Google account and the associated Google account credentials. The authentication services provided by Google are not related to the problem of proving identity of an individual hence less relevant to this thesis. The difference between the authentication services (Google acting as a Credential Service Provider) and identity proofing services (Google acting as an Identity Service Provider) is further discussed in section 6.1.3.

Using an external entity as an Identity Service Provider allows the Relying Party to have the necessary information they need to maintain some form of accountability and have *proof-of-existence* of an individual but not require actual personally identifiable information which can be used to track them or compromise their (pseudo-)anonymity. However, this raises a separate concern of the Identity Service Provider being able to track users across platforms[Vap+15] [Sun+12]. For example, if an individual uses Google as an Identity Service Provider for the

Washington Post as well as for TikTok, now Google has visibility into different Relying Parties that an individual has a digital identity established at, as well as the number of accounts that an individual has at each Relying Party. This means that Google will have access to user behavior across platforms and will be able to track what information is being used at each Relying Party.

First of all, there are concerns around the IdP (Google in this example) knowing the RP at all since the RP could be hosting a sensitive service such as chat rooms or forums around drug abuse, alcoholism, terminal diseases among others. Secondly, an IdP will probably contain verified, unique identity attributes of the user such as their full-name, phone number, government ids and even banking information. Allowing the IdP to link these attributes with an RP might be something that a user wants to avoid. Moreover, two or more RPs may collude and link user accounts on their platform to the same person based on the information provided by the IdP. Providing such a wide array of user's information to the Identity Service Provider is not ideal.

If one wants to separate this information, a mediator or broker is needed between the interactions of the Relying Party and Identity Service Provider. The broker provides a double-blind capability where the Identity Provider is not aware of who the Relying party is and vice versa. In other words, the broker allows for blind *proof-of-existence* of the users.

The Relying Party only knows that it is interacting with the broker and it is the broker that is providing *proof-of-existence*. Similarly, the Identity Service Provider only interacts with the broker and provides *proof-of-existence* to the broker. For the rest of the thesis, this broker is called the Identity Exchange (formal definition and details are provided in section 5.2.1). The Exchange ensures that neither the Relying Party nor the Identity Service Provider can track users across platforms. Moreover, the business plan and vital functionality of the Exchange is incumbent on providing trust and assurance that identity transactions mediated between Relying Party and Identity Service Provider will preserve user anonymity (further discussed in section 6.4). The Exchange ensures that Social bots can't create thousands of accounts at a platform since they won't be able to register at a legitimate Identity Service Provider and also allows for real people to create multiple accounts without the Relying Party being able to link the accounts. If the Relying Party only allows for 1 account per person (or x accounts per person), the Exchange will provide the Relying Party with the count of accounts a user has previously created with the *proof-of-existence* without disclosing the identity of the accounts

themselves. The decision to stop an individual from creating excess accounts will lie with the RP - the Exchange only providing the necessary record of number of previous accounts held by an individual.

Lastly, the question of how trust can be established between the entities involved in the identity transaction can be answered by implementing a *trust framework*. A "trust framework is a legally enforceable set of specifications, rules, and agreements that governs an identity system"[MST17] such as an Identity Federation. It is basically a set of "common operating rules" that members of a trust framework comply with. There are strict procedures for entities to be accredited in a Federation and further audits to ensure members satisfy security and privacy requirements.

4.1 Examples of Existing Brokered Identity Federations

Identity Federations and Brokered Identity Federations aren't new concepts. Splitting the tasks of identity management among trusted partners allows for streamlined identity transactions and simplifies administration of such a system as well, hence it is attractive to many organizations. Moreover, they users can pick and choose which Identity Service Provider to use and this convenience is very valuable to both the RPs and the users.

Historically, Identity Service Providers have been conjointly used for Identity Proofing and Credential Management Services. Hence, in all the examples below, the Identity Service Provider also acts as a Credential Service Provider delivering authentication services. On the other hand, the proposed architecture in this thesis co-opts this framework for the purpose of reducing fake identities in online community platforms, hence employs identity proofing services but not authentication services. More details about the architecture are provided in Chapter 5.

Brokered Identity Federations have recently become more popular and have precedents in national government schemes for the most part. They can be configured as:

- Third-party as an Identity broker.
- Third-party as a network of nodes.

4.1.1 Third-party as broker (Exchange)

Third-party as broker is a type of Identity Federation model where a proxy or a broker mediates interactions between the Relying Party and the Identity Service Provider. The third-party acts as a privacy barrier and prevents excess information being shared. *Trusted Digital Identity Framework* and *GOV.UK Verify* are two examples of such Identity Federations that exist in the wild, while *NIST Digital Identity Guidelines (800-63-3)* exist only in paper at the moment.

i - Trusted Digital Identity Framework (TDIF) in Australia

The Trusted Digital Identity Framework (TDIF) [Age20] was introduced in Australia as a move to streamline user interaction with government services. Hence, the scope of the framework is limited to federal environment and most of the members of the Identity Federation are also government entities. The purpose of the TDIF is to provide a double blind federation where Identity Service Providers do not know which Relying Party you are communicating with and vice versa. The double-blind feature is possibly due to the existence of an Identity Exchange which mediates all interactions. Double-blind allows the user to be in control of the information they share and prevent different parts of the government from being able to track users' behaviour via a large centralized population database.

The information being shared across the TDIF is highly sensitive and contains unique attributes of users, which may include their national ID number, their banking credentials, credit scores, etc. Due to involvement of such sensitive information, currently there are only two Identity Service Providers that are accredited in the federation, and one of them is owned by the government. These Identity Service Providers are DigitalID[1] (privately owned) and myGovID[2] (government owned). The digital services (Relying Parties) accessible through TDIF include: Tax file number, Unique Student Identifier, My Health Record, Youth Allowance and Newstart[3]. The current Exchange is run by Department of Human Services (DHS) [Age20]. Moreover, there are plans to establish inter-operable capabilities of TDIF-based digital identity with government digital identities in Singapore and New Zealand [DC20].

[1] https://www.digitalid.com/personal
[2] https://my.gov.au/LoginServices/main/login?execution=e1s1
[3] https://www.dta.gov.au/our-projects/digital-identity/digital-identity-system

ii - GOV.UK Verify in UK

Gov.UK Verify is very similar to the TDIF framework in Australia. However, unlike TDIF which is still in it's feedback and beta stage, Gov.UK Verify was released in 2016 and is publicly accessible right now as well. The scope of GOV.UK Verify is limited to government services and the aim is to provide an optional, user-centric, privacy-preserving digital identity for citizens. There are 7 large Identity Providers (Barclays, CitizenSafe, Digidentity, Experian, Post Office, Royal Mail, SecureIdentity) that have been certified. And there are 22 government services acting as Relying Parties accessible by GOV.UK Verify. The broker is called The Hub in UK's Federated Identity scheme and provides double-blind capability [Whi18]. The design of GOV.UK Verify was claimed to be user-centric to allow smooth transition to the service, and the Government Digital Services department announced that to create and verify account using GOV.UK Verify would only require 15 minutes [Jee16]. However, there were frequent complaints from users being unable to verify themselves.

In March of 2019, the National Audit Office in UK released the latest assessment [CG19] of GOV.UK Verify and on the ambitious predictions made by Government Digital Services (GDS) about their flagship program. The statistics paint a very morose picture of the extent of shortcomings of the digital identity scheme.

The number of people as well as government services who had taken up GOV.UK Verify was less than 20% of the early targets. In 2015, GSD predicted 25 million people would adopt this technology, but up till 2019, only 3.6 million people have been verified. Verify was expected to be self-funded by March 2018 from the profit it earned by mediating identity transactions. between the Identity Providers and the Government Services. The program was estimated to bring £873m till 2020, and these estimates were lowered to £217m in 2019 and have still missed the mark [CG19].

In 2019 the GDS announced a consultation and call for evidence[4] on the future of digital Identity in UK. After running for more than 3 years, the UK government finally publicly asked for advice on the respective roles of the private and public sectors in creating a digital identity market. The private sector and identity experts were relieved but frustrated as they believe they should have been more involved from the beginning [Gli19]. This dichotomy between expert opinion versus government direction has mired GOV.UK Verify from the very beginning.

[4]https://www.gov.uk/government/consultations/digital-identity

The government wanted minimal involvement of the profit-focused private sector when dealing with personal data of the citizens, however as has been seen by the poor take-up of GOV.UK Verify, they did not have the necessary technical expertise to pull-off such an ambitious project either.

iii - Digital Identity Guidelines (NIST 800-63-3) in US

One aspect of the National Institute of Standards and Technology (NIST) is to develop US innovation and public welfare by providing technical leadership for the country's measurement and standards infrastructure. The special publication 800-63-3 covers requirements for digital identity services implementation in the federal context only. NIST 800-63C is one of the documents in the guidelines that covers identity transactions in federated architectures - this includes both brokered and non-brokered models[GGF17].

While there is no current implementation based on the brokered identity federation requirements listen in NIST 800-63C, it acknowledges the value of federations being essential in providing privacy-enhancing communication of the public with the government digital services. The document uses similar terms for Identity Service Provider and Relying Party, but the third-party is labelled as Proxy. Since, there are no current implementations of NIST 800-63-3, there is no account for the success or failure of such a system.

4.1.2 Third party as network of nodes (Block Chain)

As opposed to a single entity as broker, a network of nodes can be used to mediate the interactions between Relying Party and Identity Service Provider. To act as an Identity Exchange some coordination is required among the nodes and block chain is one good candidate for such a purpose, with actual realizations of such a configuration existing currently. The added advantage of having a block chain is that now the information about individuals is not centralized in a database owned by the Exchange - instead it is distributed across the network of nodes and hence no one has complete visibility into the identity transactions of such a federation. This setup provides a *triple-blind* capability where the Relying Party, Identity Service Provider and even the broker (owner of the block chain) do not know who an individual is communicating with.

i - SecureKey's Verified.Me in Canada

Verified.Me was launched in 2019 by a private company SecureKey to solve the government's problem of introducing a new form of authentication that used existing credentials that users used regularly on already trusted online services. SecureKey collaborated with the top seven financial institutions in Canada (BMO, CIBC, Desjardins, National Bank of Canada, RBC, Scotiabank and TD). These financial institutions became the Verified.Me's equivilent of Identity Service Providers. Verified.Me provides access to government and private services – currently 5 services (DynaCare Plus, Equifax, FACT, Notarius and Sun Life Financial)[5]. These are the respective Relying Parties.

Canada's Verified.Me is built on top of the IBM Block chain Platform which is based on Linux Foundation's open source Hyperledger Fabric v1.2[Kir19]. HyperLedger Fabric is an implementation of distributed ledger platforms that follows specific guidelines which establishes what information on the block chain is accessible to which parties [Cac16]. Details about the block chain implementation is out of the scope of the thesis, however, the noteworthy attributes are that the triple-blind capability allows an extra layer of protection for users since now there is no central entity that knows all the information. Moreover, choosing IdPs which were highly trustworthy and covered most of the population of Canada ensured high adoption rates.

There have also been other initiatives in Canada to promote secure and private digital Identity standards. Digital ID & Authentication Council of Canada (DIACC) announced the launch of its Pan-Canadian Trust Framework (PCTF) in early Spetember 2020. PCTF defines industry standards in identity management and authentication services. Following its launch, testing of public and private sectors will begin to gauge compliance with PCTF [Paw20].

ii - Sovrin

Sovrin is a unique block chain based solution to provide a decentralized global public utility for self-sovereign identity. Self-sovereign identity (SSI) "is a term used to describe the digital movement that recognizes an individual should own and control their identity without the intervening administrative authorities"[6].

This means that when a user registers and becomes part of the *Sovrin Web of Trust*, they control

[5]https://verified.me/about/

[6]https://sovrin.org/faq/what-is-self-sovereign-identity/

information about themselves without relying on active participation from any external administrative entity. They use block chain, decentralized identifiers (DIDs) and Zero Knowledge Proofs to register, resolve, update, and revoke identity and identity claims [Sov18]. In-depth understanding of these terms and the processes involved in Sovrin are not relevant to this thesis. Since this is a completely new system, terms like Identity Service Providers, Relying Parties and Exchanges are not applicable to the system - instead Issuer, Owner and Verifier are terms used.

Sovrin should not even be considered part of Identity Federations since it does not rely on digital identity being "issued" by an Identity Provider - which is fundamental to Federation systems.

4.2 Looking ahead

This chapter introduces the concept of blind *proof-of-existence* and how such an assertion can be provided using a brokered identity federation. Furthermore, two configuration of existing architecture of brokered federation are are discussed, one where the broker is an entity, and another where the broker is a block chain.

While using a block chain adds an additional layer of security by providing triple-blind capability where user activity is not tracked even by the owner of the block chain service, it is not the approach recommended by this thesis. The largest obstacle for a global level block chain based identity system is slow adoption - by users as well as online services which would act as Relying Parties or Identity Service Providers. Parties involved would need to build completely new configurations to comply with a block chain based solution so it has a higher on-boarding cost. Hence, even though it is technically viable solution, it is not the most attractive solution.

On the other hand, using a third-party as a broker in an identity federation does not require much changes from both the Identity Service Provider and Relying Party. Moreover, since this thesis is using the brokered identity federation for only identity proofing purposes, that means that interaction with the Identity Provider is a one-time interaction for one account on a Relying Party (further discussed in section 5.2.2 and also in section 6.1.3). So, there is not a lot of burden on the user either.

The next chapters will provide in-depth information and recommendations on how to realize

an architecture which can be used to limit social bots on a global scale while preserving user anonymity.

Chapter 5

Digital Identity Framework in a Brokered Federation

5.1 Architecture Overview

The proposed architecture follows a federated model of identity. Specifically a brokered federation where identity transactions across a networked system are intermediated by a third-party, also sometimes referred to as a Federation proxy or an Exchange. The Exchange provides a privacy barrier between the two main parties of the Identity Federations: the Relying Parties and the Identity Service Providers (in depth definitions can be found in section: Key Entities). The goal of the privacy barrier is to limit the identity information shared by the parties across the Exchange and protect end-users from unnecessary data sharing and data leakage that could be used to track them.

The architecture includes a set of standards and compliance requirements that all participants of the federation must follow through on. User consent and visibility of identity attributes (if any) is centralized at the Exchange. Identity Service Providers are responsible for identity proofing, identity verification and identity management and must pass the compliance requirements of associated accreditation authority. For a federated identity model to be effective, it needs to support a diverse set of technical protocols and hence the architecture includes a technical integration standard for interactions between participants of a federation.

The architecture of this Brokered Federation is heavily influenced by Proxied Federation in NIST Digital Identity Guidelines (800-63-3)[GGF17], Australia's Trusted Digital Identity Framework (TDIF)[Age20] and UK's GOV.UK Verify[Ser] [Whi18]. NIST 800-63-3 documents are the current guidelines for federal agencies implementing digital identity services and provide a more holistic overview of identity transactions, within and outside the purview of federations. On the other hand, Australia's TDIF and UK's Verify are specifically targeted at making government service available online while maintaining a double-blind privacy requirements. Our proposed architecture differs from these frameworks in three major ways which will be discussed in more detail later on in the chapter. These are:

- **Re-purposed.** Brokered Identity Federation's have never before been used to curb the problem of mass fake account creation on online community platforms, to the best of the author's knowledge.
- **Global.** Instead of accessing government services or being limited to the federal environment, the current proposal makes a case for global adoption for any commercial (or otherwise) Relying Party that wants to use its services.
- **One-time Assertion.** Previous architectures use the identity federation to provide repeated-access capabilities, i.e. the IdP acts as a Credential Service Provider as well. This means that a user would use their IdP every single time they logged into the RP of their choice, through the brokered federation, and they could use this one credential to log into multiple RPs. The current proposal does not require such capability, instead it focuses on using IdP to provide a One-Time assertion to the RP when a user is creating an account on the RP, as proof that the user has an account on the IdP. The proof provides the RP with a certain level of assurance of the existence of the user as a 'natural person' and/or the number of times a subscriber has used the IdP to create an account on that particular RP.

These main takeaways are further discussed in Section 6.1 after the architecture description is laid out in this chapter.

5.2 Key Concepts

5.2.1 Key Entities

The key entities in the proposed Identity Federation are:

i - Relying Party (RP)

Relying Parties, sometimes called Service Providers, are the entities that a user is trying to access. They are digital services that rely on assertions of user's identity provided by an Identity Service Provider through the privacy-preserving layer of an Exchange.

To enable access of a user to their digital services, a Relying Party needs:

- A method of authentication. A user needs to provide digital credentials (such as Username and Password) to access the digital services.
- Verified Identity Attributes. The digital service may need to have access to verified attributes such as name, email id or just proof of a name existing without knowing the name itself.

Moreover, the Relying Party needs to trust the authentication as well as the verified identity attributes. This trust is provided by the Level of Assurance for identity and attributes. Relying Parties can set the level of assurance required to access their digital services and allows flexibility in terms of what identity transactions they deem acceptable.

It is important to define what a Relying Party means in the context of the federation. It is possible that one organization entity provides multiple digital services and considers all of them as separate Relying Parties with separate authentication and attribute verification requirements, such as how Instagram and Whats App are both owned by Facebook by for all user intents and purposes, they are separate entities (hence two different Relying Parties). On the other hand, it is possible that multiple organization entities have an interlinked platform and are considered a single Relying Party, such as how Gmail and YouTube do not need separate authentication, neither does any other application on Gsuite, as long as you login to one - so for the purposes of our architecture, they will be considered one Relying Party. This consideration is important because it (a) impacts the identity linkages managed by the Exchange - all digital services that make up one Relying Party in context of the federation will

have one identity, and (b) impacts how user consent is obtained - the user must know all the organization entities that their information will be shared with through the Exchange.

ii - Identity Service Provider (IdP)

Defining identity, just like defining accountability and anonymity has been difficult in the online space. There exist many different definitions for identity and what it entails - is it a unique attribute that a natural person possesses, is it an unchangeable attribute or is it a set of (possibly) non-unique attributes that define a user's identity? Chapter 3 provides some discussion on what online identity entails.

For the purpose of this architecture, identity is a collection of attributes that represents an individual. It is context specific, so a user's identity can be different on different platforms and there are allowances in the framework for the RP to define what attributes represent a natural person. Hence, an Identity Service Provider (IdP) is an accredited service that verifies identity attributes, manages the verified attributes and binds these attributes to a credential which can be used as an authoritative source of the identity attributes of a natural person. In addition, an IdP must provide some level of auditability services to get accredited to the Federation and maintain trust of its service.

Many different architectures break down the role of IdP into various identity related services. For example, the NIST Digital Identity Guidelines, primarily uses the term Credential Service Provider (CSP) to describe the service of issuing credentials and registering authenticators to verify identity to a RP. It only uses the term "Identity Provider" as separate from CSP when it describes the federated scenario since in the general Federated environment, the "Identity Provider" may or may not provide the services associated with the CSP. However, that is not the case for my proposed architecture. Hence, there is no need to divide the identity related services to multiple different entities. For a more detailed comparison (or mapping) of the terminologies for these schemes, refer to Exhibit A-1 in the Appendix A.

iii - Identity Exchange (IX)

The Identity Exchange is the main component that makes brokered federation possible. It is an entity that manages the identity transactions, flow of identity attributes and assertions as well as requests of identity attributes and assertions, between the participants of the federa-

tion.

The Exchange allows for easy technical integration between Relying Parties (RP) and Identity Service Providers (IdP), it provides double-blind ability that fulfills privacy and anonymity requirements of users, IdPs and RPs and it is a central point for user interaction with the Federation. Details of Exchange's functionality are provided in section 5.2.2.

The proposed architecture does not require that only one exchange exist in a federation. Having the option of provision of multiple Exchanges allows for **Global** reach of the identity system. However, it would be preferable for the market to not be saturated by too many Exchanges as it might complicate the user experience and hence reduce the adoption-level of such a Federation Scheme.

iv -User

The user that benefits from the federation is not an accredited member of the identity system itself. Instead, it is an individual who interacts with the system from the outside, with the purpose of obtaining a service from the RP. A *user agent* is the browser or the operating system that the user employs for such interactions and there is no need in this architecture, for the user agent to remain constant among different interactions.

5.2.2 Brokered Identity Federation and Identity Mappings

A key concept in implementing Brokered Identity Federation is the identity mapping within the Exchange that realizes a privacy-preserving accountability mechanism. This subsection first distinguishes terminology of Identity Federation and Single-Sign On, and then proceeds with explanation of one-time assertions and pseudonymous identity mappings.

Identity Federation v. SSO

There is sometimes confusion between the terms *Identity Federation* and *Single-Sign On (SSO)*, and sometimes these terms are used interchangeably. For the purpose of this thesis, these are two very separate terms and SSO is not relevant to the proposed architecture. SSO is a tool which allows a user to use the same credentials (usually a username and password) to access multiple websites *within* an organization. For example, your school might have a different website for the library, a different one for accessing courses and a different one for

your grades but they might all be accessible through the same credentials - this prevents a student from needing to learn multiple passwords and allows streamlining of the databases on back-end as well. On the other hand, Federated Identity is a model that has agreements and standards which allows some entities to take the role of identity providers and some become identity consumers. This distributed arrangement provides functions to share identities between *different* organizations. While SSO requires only a single set of credentials which allows access to multiple applications of one organization, Federated Identity maps across multiple organizations and user identity authorization is provided through specific standards (OpenID and SAML being the most popular) to establish trust relationships.

One-time Assertion and Identity Mapping

For the proposed architecture, SSO is not even necessary for the functionality of the identity federation. The purpose of the federation is to provide a one-time assertion of a user's identity from an approved Identity Service Provider to a new Relying Party. The assertion is provided once since it is only required when a user makes a new account on an RP. If a user has an already existing account on the RP when the brokered federation is introduced, the RP can still prompt the user to go through the **one-time assertion**. This one-time assertion is the *proof-of-existence* discussed in chapter 4 which is used to establish that the user asking to create an account on a RP is a real person and also provides the count of the accounts a person has created on a specific RP using a specific IdP.

To prevent linkability and provide privacy-preserving interactions, the proposed architecture provides a broker model where all interactions are mediated by the Identity Exchange. This double-blind setting where the RP does not know the identity of the IdP and vice versa limits tracking and profiling of users across the many services they access.

The Identity Exchange keeps pseudonymous unique identity mappings that allow the Exchange to keep track of the number of times a user has used the same IdP to create multiple accounts on a RP or across multiple RPs. These persistent linkages also provide an audit trail without compromising on the privacy preserving nature of the identity federation.

The diagram in 5-1 shows an example of identity mapping of a user Amna. Amna has 1 account on RP_B and 2 accounts on RP_A. There are multiple legitimate reasons why a user would want to have multiple accounts on a platform such as starting an account for their child, having

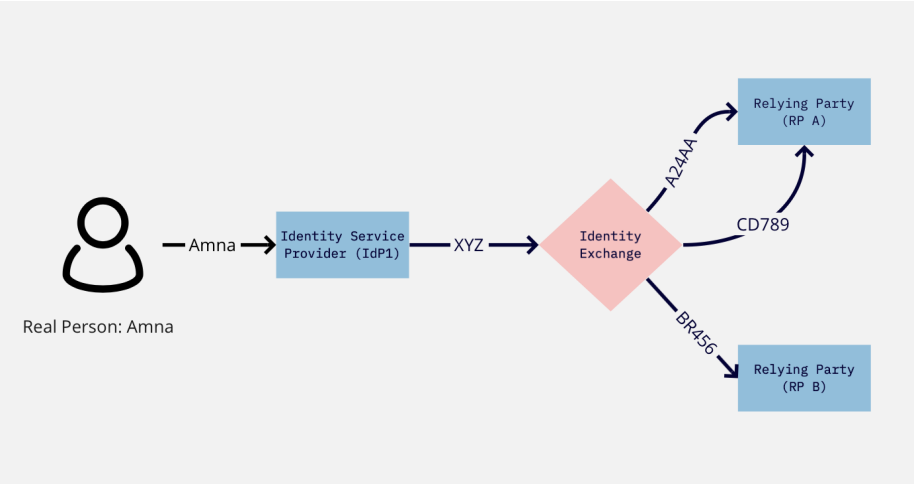

Figure 5-1: Mapping of a User's Identity across an Identity Exchange

an account for their business, etc.

Some important features:

- For the same user at an IdP, the same identity mapping will be created every single time. So, Amna on IdP1 will always generate XYZ.
- The same IdP can be used to create multiple accounts on the RP (if the RP allows having multiple accounts to a certain limit). The RP will not be able to identify which accounts belong to the same person as the identity mapping from the Exchange produces different values. In the case of Amna, RP_A does not know that identity A24AA and CD789 are the same person. However, this feature is only possible for RPs that don't ask for any verified (and unique) identity attributes from the user through the IdP via the Exchange. If the RP asks for Amna's verified Name, DOB and country of residence each time she creates an account, the RP will be able to connect both the accounts to the same person.
- Even though RP_A does not know A24AA and CD789 are the same person. RP_A does get the count of previous accounts a natural person has created on this platform. The first time Amna creates an account, The IX will inform RP_A that this is the first account created by a user. When Amna creates a second account, the IX will inform RP_A that CD789 has a previous account on their platform. Based on the rules established by the RP, they can decide whether they will allow this second account or not.

RP Name	RP link	IdP link	IdP Name
RP_A	A24AA, CD789	XYZ*	IdP_1
RP_B	BR456	XYZ*	IdP_1

* The asterisk represents that not the actual value but the hash of the value is stored.

Figure 5-2: Mapping of a User's Identity in an Identity Exchange Storage View

The table in 5-2 shows the information that the Exchange will have based on Amna's interactions in diagram 5-1. These identity mappings will ensure that no unnecessary information is leaked from the IdP to the RP, and no correlations can be generated by two RPs colluding together.

The information that must be stored in the Exchange is the name of the RP and IdP, and the associated RP link and IdP link. The design choice to generate and store an *IdP link* (one way hash of the user identifier on the IdP) and not the user identifier itself is to ensure that no irreparable information is exposed if the Exchange experiences a data breach. A malicious entity will not be able to provide a stolen IdP link to establish new fraudulent identities on RPs as the IdP link itself is not the information sent over from the IdP, hence it is useless even if leaked.

The design choice to generate and store an *RP link* (one way hash of the random identifier an RP sends when making a request to the Exchange) as opposed to just the count of accounts on an RP asserted by the IdP is to allow for the option of a more robust audit trail. If an RP complains that a set of users verified by the Exchange were fake accounts, the Exchange will have the option to trace back the RP links to the IdP link and block the IdP from creating further accounts.

5.2.3 Levels of Assurances

Another feature of the architecture is to provide different levels of assurances for different identity functions instead of expecting all participants to adhere to one gold standard of

identity requirements. The major three identity functions are identity proofing (by the IdP), authentication (by the IdP) and federation processes (by the Exchange). Table 5.1 shows the description of what each assurance type means, and Exhibit A.2 in the appendix provides detail of the levels of varying strengths possible for each type of assurance.

For example, an RP can request IAL2, AAL3 and FAL2. Only the IdPs that are accredited to IAL2 (or higher) and AAL3 will be displayed to users as the options even if other IdPs are part of the federation. Moreover, the Federation assurance level of FAL2 means that the Exchange involved can cater to the assurance level requirement of the RP. An RP will choose different levels of assurance based on what identity information they request and what would be the impact if a bad actor is able to commit fraud.

The advantage of allowing for different strengths and levels of assurances means that more number of of IdPs can be qualified to be part of the brokered identity federation. This results in more choice for user, higher accessibility of the entire system and appropriate risk-based approach to identity transactions.

Assurance Type	Description	Possible levels[1]
Identity Assurance Level (IAL)	Identity Assurance Levels are used in the context of Identity Proofing, a process to validate the correctness of the identity attributes of a natural person and verify that the attributes belong to the natural subject.	IAL1, IAL2, IAL3 and IAL4
Authenticator Assurance Level (AAL)	Authentication Assurance Levels (AALs) are used to describe the strength of the credential used to authenticate.	AAL1, AAL2 and AAL3
Federation Assurance Level (FAL)	Federation Assurance Level (FAL) describes aspects of the assertion and federation protocol used in a given transaction.	FAL1, FAL2 and FAL3

Table 5.1: Levels of Assurances for identity proofing, authentication and federation

5.2.4 Key Interaction

Figure 5-3 shows the main interactions between a user and the Identity Federation. Flow A in the diagram shows the key steps a user takes to create an account on an RP in the federation. These are:

[1]higher number means more stringent requirements.

1. User tries to create an account on the RP. The RP redirects user to the Identity Exchange as part of the sign up process.
2. On the Identity Exchange, the user selects an IdP from a range of IdPs that satisfy RP's Level of Assurance requirements.
3. After the user is redirected to their chosen IdP, they attempt to access the platform. If the user already has an account on the IdP, they simply authenticate themselves with their credentials. Otherwise,the user creates a new account on the IdP and verifies their identity (and identity attributes) through whichever Identity-proofing mechanism the IdP employs.
4. The user then gives consent for share attributes (proof of attributes) to the Exchange.
5. From the non-user(entity) perspective, the next step is the identity mapping between the identifiers of the requesting RP and IdP. However, the user is unaware of this step.
6. From the user's perspective, the Exchange again asks for user consent and then shares the information with the RP. The RP could have requested for proof of identity, identity attributes like name, DOB and email, Levels of Assurance from the IdP and the Federation and count of previous accounts on the RP from the IDPowner (user).
7. User is returned to RP and continues to create an account.

Flow B shows the steps a user takes to re-authenticate into the same account on the RP. As flow B shows, authenticating into the account on the RP does not need any interaction with the Identity Federation and all future interactions will also be with only the RP itself as well.

5.3 Protocol Support

The current standard for protocols that support identity federations are OpenID Connect 1.0 (OIDC 1.0) and Security Assertion Markup Language 2.0 (SAML 2.0). These two protocols represent the most commonly used technologies for identity transactions in identity federations, but they are not an exhaustive list. The purpose of both SAML and OIDC is the same in identity federations, the difference lies in the underlying mechanisms for each. Understanding these mechanisms is out of scope for the thesis.

OIDC is relatively a new protocol in comparison to SAML, hence SAML has higher existing adoption but OIDC is attractive since it is easier to use than SAML. The thesis does not

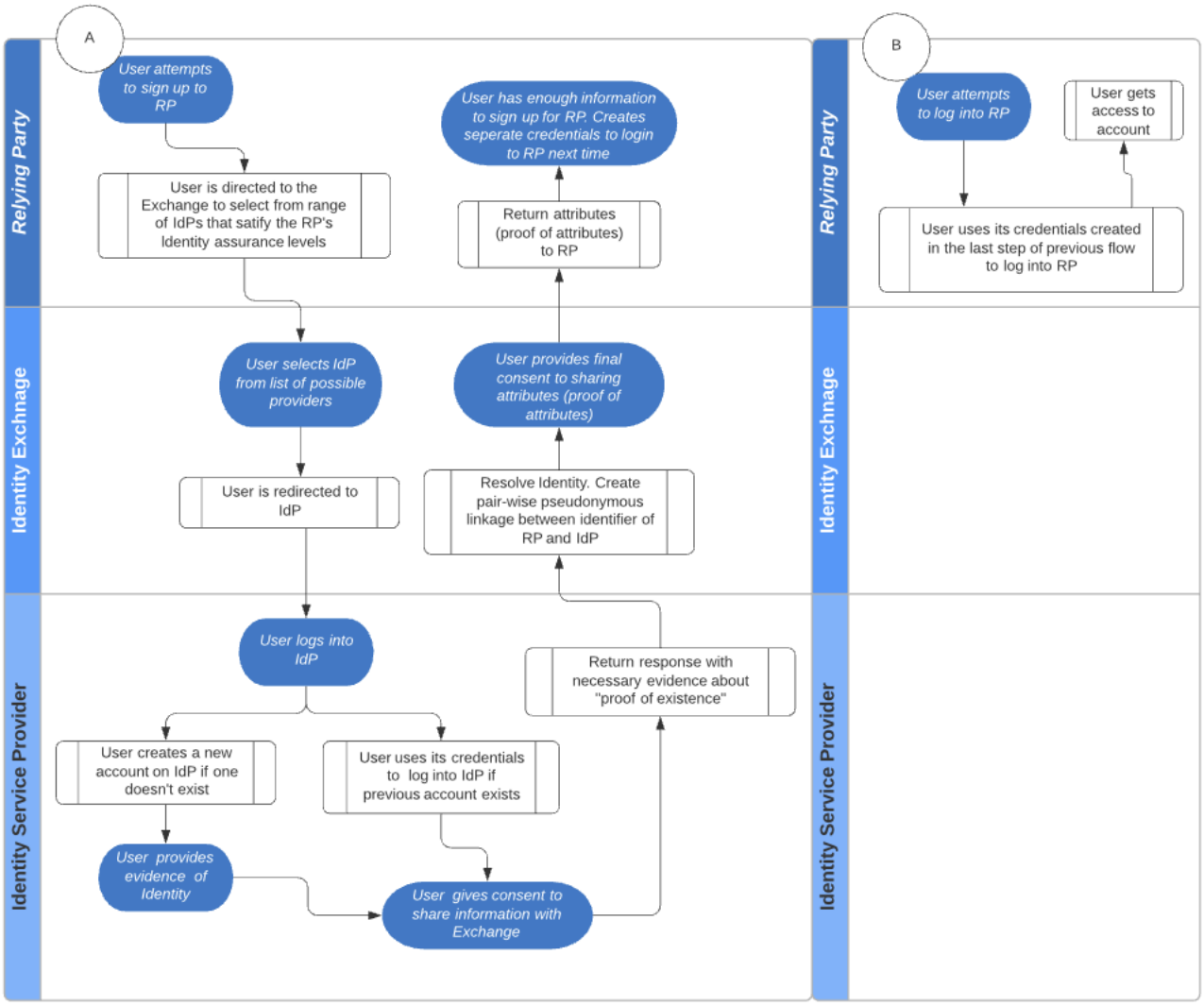

Figure 5-3: Major User Interaction with the system

recommend any preference to either protocol. However, the Exchange as the trusted broker must support both integrations and the RP and IdP can choose whichever protocol integration they want to implement or have already implemented. A detailed example of the the key interactions while implementing OpenID Connect can be found in the Appendix B.

5.4 Security Requirements

Any interaction online is susceptible to attacks and threats carried out by malicious actors. In a brokered identity federation, where the identity transactions require more coordination, as they go from the IdP, through the Exchange, and to the RP, there are additional opportunities for attackers to exploit possible vulnerabilities.

Attackers who want to thwart this architecture would want to intercept assertions, i.e. *proof-of-existence*, from the the IdP and re-use them to impersonate an individual. The attacker could be a new entity or it could be a misbehaving IdP, RP or a user of either an RP or IdP. To prevent such attacks and ensure the information passed through the Exchange maintains its integrity and confidentiality, there is a need to establish preventative measures to ensure the system is above reproach. Common attacks on an identify federation and the accompanying mitigation strategies are shown in Exhibit A.3 in the Appendix.

Apart from external threats, there is also a need to regulate how the Exchange handles the users' functional data internally. Due to the Exchange's central position in the federation, it has an important role in auditing and logging, while maintaining privacy requirements of minimal data storage (discussed further in section 5.5).

5.4.1 Auditing

The Identity Exchange is the only party in the federation that has visibility over all the interactions between RP and IdP. Not only is the Exchange the only entity that knows the identity of the RP and the IdP, but also has visibility on any attributes (in addition to *proof-of-existence*) shared by the IdP.

For auditing purposes, the Exchange must provide a historical record of all identity transactions, successful and otherwise. This encompasses all requests and responses between (i) IdP and Exchange, and (ii) Exchange and RP. The information logged can vary based on the

standard set by the Exchange or it can be individually negotiated with the parties (RP and IdP) involved.

For most intents and purposes, the Exchange only needs to mediate interactions of *proof-of-existence*, however some RPs might request attributes such as name, date of birth or whether the user is above 18 or not. For such purposes, the Exchange must retain the name of the identity attribute that was shared with the RP, but not the value of the attribute itself. The attribute value provides no benefit for the audit history, however it would be excess (and unnecessary) information stored by the Exchange. Hence, values of the identity attributes will not be retained by the Exchange. The information logged must only be limited to that required to complete audit history and maintain trust in the federation. The audit history may include (i) metadata information such as timestamp, protocol used, any cryptographic keys necessary, (ii) Consent information such as consent decision, duration of consent and (iii) major interaction information such as RP name, IdP name, RP link, IdP link and any assertions or attribute names.

5.5 Privacy Requirements

To ensure user anonymity on RPs and maintaining privacy-preserving identity interactions through the Exchange, it is important that the architecture specify privacy requirements and recommendations. The major privacy considerations are providing unlinkability between RPs and IdPs, as well as among different RPs, minimizing identity data stored by the Exchange as well data transferred to the RP, ensure explicit and informed user consent and follow internationally established privacy guidelines.

5.5.1 Limit Tracking and Linkability

Brokered Identity Federation has the capability to limit direct interactions between RP and IdP which makes them a good candidate to establish a privacy-preserving identity transaction environment. As discussed in section 5.2.2, identity mapping withing the Exchange limits the privacy risks of user tracking and profiling. For example, if an IdP interacted directly with multiple RPs, the IdP would be able to create a profile which could describe user behaviour and this information could be used for advertising purposes. Users might have concerns about entering an Identity Federation which allows for such non-identity related attribute collection

to occur.

The Exchange will take measures to ensure that any processing of identity attributes or identity assertions occur without association to individuals, their devices and their IdPs beyond the necessary requirement of the identity transaction. This dissociation is enabled by pseudonymous unique identity mappings in the Exchange and storage of only limited data in the long-term.

5.5.2 Data Minimization

The proposed architecture aims to reduce the amount of information available to the RP. Firstly, the RP does not know the identity of the IdP - only that the IdP follows the Levels of Assurance requirement set by the RP (details in 5.2.3). Secondly, even though the IdP in the federation might provide extra information to the Exchange, the Exchange will ensure that no additional attributes beyond what the RP requested is transmitted to them.

Moreover, there are scenarios where the RP does not need the identity attribute, only a reference to it. In such scenarios the RP must request the reference and not the actual value of the identity attribute. For example, the most common scenario for our architecture is that the RP must know that the person is real so just the assertion of existence of legal name is enough rather than what the user's actual legal name is. Similarly, sometimes and RP only needs to know that a user is above the age of 18, in which case they do not actually need to know the birth date of the user just Boolean response is enough. This limits the RP's collection of unnecessary PII.

5.5.3 User Consent

The user must give express consent before any information is shared with the RP. Consent will be first asked when the user logs into their IdP and information needs to be shared with the Exchange. The second time consent is required is when the information is being shared with the RP. Consent decisions may be saved by the IdP or the Exchange so they don't pop up every time the user engages with an Exchange or a particular IdP.

The notice of information being collected and the identity attributes being shared must be presented to the user in a readable and clear format so that the user is fully aware of the information being passed by the Exchange. If there are any optional attributes, the user must

be clearly presented with this option as well as the ability to decline the information being shared with the RP completely.

Following recomended consent guidelines ensures that users make informed decisions about their data. While these guidelines might seem straightforward, presenting information that is easy to understand, informative and does not impair user experience is quite a complex problem. There is a vast amount of literature that covers why privacy by informed user consent is such a difficult topic. For example, there are discussions on the readability of consent notices [LMR13], how small design choices have large impacts on user interaction with consent notices[Utz+19] and the possibilities of negotiating different access to user's personal information [Baa+15].

Moreover, the thesis recommends that users should be given more authority over the data that has been shared with the Exchange. While this is not critical to the infrastructure, it is highly recommended to establish trust in the system. This authority over the data can come in the form the user's ability to revoke consent and request for the "Right to be Forgotten". For such scenarios, the Exchange must create a dashboard of the information stored about a user in the context of a particular IdP and provide the option for it to be deleted. The architecture proposal does not specify whether the RPs who used such information will be informed when such revocation occurs. However, it could be possible for RPs to decide whether they want user revocation information or not. They can also then choose whether they need a user who has revoked their consent and deleted their information need to provide a new *proof-of-existence*.

5.5.4 Privacy Compliance and Governance

For IdPs to join the Brokered Identity Federation, they must go through an accreditation process which ensures they comply with the security and privacy guidelines (further details in section 5.7). This accreditation process ensures that entities commission a Privacy Impact Assessment (PIA) to review the privacy impacts of the services offered by the IdP. As part of the accreditation process, IdPs will go through privacy audits regularly to ensure they are inline with the current requirements.

Similarly, all members of the federation must follow the European Union General Data Protection Regulations (EU GDPR) to ensure data use risks are minimized.

5.6 Usability Requirements

Usability is one of the most important considerations when it comes to user adoption for any system that requires user engagement. Usability comes under the purview of User Experience and is defined as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"[2]. From the user perspective, the Brokered Identity Federation should not encumber users too much.

While approaches and usability considerations of proving a user's identity have been well-researched, there is a dearth of conclusive literature about identity proofing in the context of a double-blind Brokered Identity Federation. Research on the users' perspective on trust and benefits of the brokered model as well their mental models regarding how the process occurs is at its nascent stage.

5.6.1 User Perspectives

Users may have different mental models of what Online Identity represents. It is important to convey what identity data means in the context of the brokered model. There is a need to establish reliable expectations for what information is being used, how it is being transferred and what entity knows what information.

A typical user would not be familiar with terms like 'RP' or 'IdP', hence they cannot be expected to understand how the existence of these entities and the Exchange prevent tracking and profiling. There is a need to build user *trust* in the system and provide an understanding for the benefits as well as the risks. While an engineer or designer might think of this *trust* being built on notions of cryptography, OpenID Connect, levels of assurance etc., the user does not have the knowledge of such technical jargon nor is the user expected to have such information.

Some of the factors that must be considered to clearly establish identity and how it is being used in this context are as follows:

- Clearly establish the role of each entity in the federation that is unique, meaningful, and descriptive.

[2]https://www.iso.org/standard/77520.html

- Provide information about data ownership and usage to the users. Include information on what information is long-term v. transient.
  - An effective method to display to users what information is being held by the Exchange in the long-term is to provide a Dashboard where a user can see what RPs requested information from an IdP through the Exchange and provide meta information such as time stamps, success/failure rates well as what identity attributes were shared.
- Allow the users to have the ability to easily verify, view, and update attributes on the IdP.
  - Updating attributes on the IdP should have no impact on previous or future interactions with the Exchange.
- Users should also be able to delete their identity. Removing all information from an Exchange, including the history of transactions, should be possible.
  - There is a need to consider the resultant audit, legal, or policy constraints that such an action might cause.
- Provide users with easy access to notice and consent forms as well as privacy policy documentation which is easy-to-read and understand.
- Provide users with the mean to validate the complete separation between RP and IdP. Complete understanding of the transaction flow among the entities is vital for the user to be able to trust the double-blind feature provided by the Exchange.
- Minimize user actions and steps required. For example, the ability to remember a user consent decision for a particular IdP or a particular RP.
- Reduce superfluous information that might confuse the user. For example, the user has no need to understand what an IdP means or what is the value of their pseudonymous identifier at the Exchange.

5.6.2 Key Performance Indicators

Measuring user satisfaction is the best move to continually improve how the user interacts with the Exchange, the IdP and the RP in the context of identity transactions. UK's brokered federation GOV.UK Verify (discussed in section 4.1.1) has a detailed guideline to measure

success[3].

One of the suggested ways to track user satisfaction is to request feedback from users. These requests could be (i) passive where just the option of providing feedback is available and easily accessible for the user, (ii) presented at the end of the user transaction by the RP or (iii) prompted for when a user drops-out of the service or revokes access of the Exchange. Feedback provides useful information too improve how the federation is operated and what are pain points for the users.

However, getting direct feedback from users on aspects that are not the primary function of the RP (the service they want to access) is usually ineffective. Rarely would a user be interested in providing feedback on how identification or authentication process were perceived, unless there is an attached reward for completing feedback.

Hence, performance indicators that don't require direct user actions are preferred. One such performance analysis tool is measuring *completion rate* of the identity transactions. The steps involved are:

1. Count the number of transactions that were completed, i.e the count of identity transactions where the request of the RP was fulfilled.
2. Count the total number of transactions (includes partial and failed interactions).
3. Divide step 1 by step 2 and show the result as a percentage

Measuring performance indicators is a good practice as a largely positive user satisfaction rate might trigger an increase in demand for the brokered federation model which might cause an increase in IdPs and RPs wanting to be involved.

5.7 Accreditation Process

For the Architecture to be sustainable, there is a need to maintain certain levels of privacy and security requirements. Before an IdP is on-boarded to an Identity Federation, they must follow through on certain requirements to ensure their identity services are up to par. This is a critical step as the trust in the Exchange and the system is based on how accurately the Exchange mediates identity transactions. If there is any doubt that the Exchange is provided

[3]https://www.gov.uk/service-manual/measuring-success/measuring-user-satisfaction

faulty or fraudulent information, RPs will no longer be willing to be part of the identity federation.

The applicant IdP should provide proof of the service complying with the privacy guidelines, security requirements and usability criteria to be successfully on-boarded. Then, there must be yearly evaluations to keep up with upcoming cyber-threats to maintain accreditation.

IdPs may need to update their Identity service to be compatible with OIDC and SAML protocols. There may also be a need to use third-party assessors to verify complinace with security and privacy guidelines such as Privacy Impact Assessment (PIA) or NIST's Security and Privacy Controls for Information Systems and Organizations (NIST 800-53)[4].

[4]https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 6

Other Considerations

6.1 Existing Brokered Federations v. Proposed Architecture

Brokered Identity Federation is a relatively new approach to carrying out privacy-preserving identity transactions. Currently, it is primarily used in National ID schemes with examples provided in Chapter 4. Such federations are restrictive in the sense that they are overseen by a government entity and typically provide access to RPs which are government services. The fundamental use case of these National brokered models is to prevent mass government surveillance which can be possible if a user's tax information, credit information, healthcare, etc are all tracked and linked.

The proposed architecture uses the same building blocks as the brokered federations that exist in the wild but addresses a different problem, with a different scope and hence a different use case.

6.1.1 Re-purpose Exchange to limit fake IDs and bots

We have established that Brokered Federations provide double-blind capabilities that prevent IdPs and RPs from tracking users and creating behavioural profiles. The need to have some evidence of a user's real world existence has become more and more important as online community platforms have become attributed to racial violence, impacting US (and global) elections and unwittingly compromising democracies. These platforms are being abused by

social botnets by creating thousands of fake accounts to promote propaganda and spread disinformation. Increasing *effective* accountability measures have become necessary. Brokered federation models applied to a global scale can ameliorate the disastrous consequences of disinformation spread by fake social bots without compromising on user anonymity and user experience.

Instead of preventing government surveillance while providing highly sensitive PII, the proposed model re-purposes the Exchange to prevent user tracking by RPs and IdPs while providing only the necessary *proof-of-existence* to thwart fake accounts.

6.1.2 Provision for multiple Exchanges - Global Reach

The scope of the proposed identity federation is not limited to national boundaries or the federal environment, which has been the case for previous schemes which have gained repute. The current proposal makes a case for global adoption for any commercial (or otherwise) Relying Party that wants to use its services.

For global coverage, it is expected that competition may arise in the form of multiple federations, each with their own Exchange or more than one Exchange wanting to collaborate and becoming part of one Identity Federation. The architecture does not oppose the existence and growth of multiple Exchanges as the aim is to ensure that a person can prove their existence, hence leeway exists for real people to create multiple accounts on a RP by either not exceeding the account limit set by the RP, using multiple IdPs or using different Exchanges. This leeway is not an abuse of the system as the threat is not a user creating 5 or 10 accounts, but an individual creating thousands of accounts on a RP. The aim is to have additional cost of proving an individual's existence which mitigates the harm caused by malicious actors controlling large social botnets.

6.1.3 One-time Assertion

Section 5.2.2 in Digital Identity Framework in a Brokered Federation chapter highlights how RPs only require a one-time *proof-of-existence* of a new user when creating an account on a platform and this one-time assertion can be provided by any IdP within a federation that fulfills the level of assurance requirements set by the RP. An example of this one-time assertion with OpenID Connect protocol can be found in the Appendix B.

This is the most substantial difference in the current proposal in comparison to other brokered identity federations. Not only is the IdP used only *once* for an account at a RP, but there is no expected need to transmit highly sensitive PII across the Exchange because a RP which is an online community platform does not need information such as a user's SSN number, or credit score or house address. In earlier architectures, an IdP not only provides identity proofing services but also acts as a Credential Service Provider (CSP) and provides authentication services. In Australia's Trusted Digital Identity Framework (TDIF) [Age20], "Credential Service Providers generate, bind and distribute Credentials to individuals or can include the binding and management of Credentials generated by individuals". This means that the credentials used to access the IdP (acting as the CSP in this instance), can authenticate a user to a RP as well.

The proposed architecture does not require the IdP to take on the role of the CSP as only a one-time assertion or *proof-of-existence* is necessary. The proof provides the RP with a certain level of assurance of the existence of the user as a 'natural person' and/or the number of times a user has used the IdP to create an account on that particular RP. Using the same CSP to gain access to a RP every single time might add additional burden on the user and increase the risk of information being tracked by the RP. For example, if a user is only using an IdP for only one RP, the IdP might be able to track user behaviour despite not knowing what RP is being accessed because the user will have to use the IdP every time they try to gain access to the specific RP.

6.2 How to motivate users to enroll in such a scheme?

When a RP becomes part of the proposed brokered Identity Federation, they have to choose how to deal with users and the provision of *proof-of-existence*. Their choices are:

- **Compulsory**. Each and every user will have to conform to the new policy within a certain time period to provide a *proof-of-existence* through the Exchange, whether new or old.
- **Optional**. The platform can keep it optional for users to provide this *proof-of-existence* and aid with making the platform more secure against fake accounts.

The first option seems very stringent and might make it hard for the users to comply with.

Figure 6-1: Similar looking Trump accounts on Twitter but only the left-most is a verified account with the blue badge check mark next to Trump's name. [2]

But in the end it is the choice of the RP on how they want to roll-out this scheme and will not affect any of the functionality of the Exchange. A RP will make such a choice based on what type of service they are running and what are reasonable expectations from users.

The second option provides a lot of leeway for the users but keeping account verification through the Exchange optional means that the adoptability level could be very low and prove to be ineffective then. It is important to incentivize users to go through this identity proofing. While this thesis does not explore in depth usability constraints, but future work might provide valuable insights in this matter.

One possible option is to give users who go through the identity proofing step a higher status or tier on the platform. As of right now, there are two classes of users on most platforms: "verified" and "unverified". Verified accounts are usually celebrity or important figure accounts that are high targets for fake accounts. Figure 6-1 shows a real Donald Trump account next to 3 fake ones, and only the small blue check mark next to the name provides information that the account is verified to actually be owned by the actual person, Donald Trump[1]. If a user who provides their blind *proof-of-existence* gets a badge similar to "verified", it would add a level of credibility to them as well. Such a badge could be labelled as "real user" (so that it is different from "verified"), with the aim that their comments or posts would be deemed more trustworthy than the "unverified" crowd.

[1]https://www.computerhope.com/issues/ch001850.htm

[2]As of 8 January, 2021, @RealDonaldTrump's account was suspended for violating Twitter's Glorification of

6.3 Who should be the IdPs?

The range of IdPs that a user will have to choose from is a very critical consideration for the Brokered Federation to be successfully adopted. The range of IdPs will decide the accessibility of the overall Federation. One of the reasons for Canada's success of Brokered Identity Model for their government ID was that they chose banks as their IdPs which covered a large proportion of their population (discussed in section 4.1.2) and hence it was easily accessible to everyone.

For a global identity federation, this accessibility constraint becomes even more vital. The federation might choose to take up various government eIDs and global banks as their IdPs but there is a high probability they will miss a large portion of users. For this purpose, the federation might turn to established global Identity Providers who traditionally have had a customer base of financial services and trade services. Trulio[3] is such an identity verification service that does id and document verification. It leverages 400+ trusted global data sources across 195+ countries that includes credit bureaus, electoral rolls, national IDs, mobile network operators, etc. Getting services like Trulio on board, will allow accessibility to become much easier. Trulio is not alone in terms of such services rendered. Kantara[4], Veriff[5], OneLogin[6] and Au10tix[7] are just some services that can fill similar roles.

However, the possibility of corner cases that are not catered by such IdPs is still possible and it is important to take them into account to when implementing such a system. For example, minors, disenfranchised citizens, un-banked people, political dissidents and people with limited technological infrastructure are some of the classes of people that will need to have special consideration for the scheme to become truly global.

6.4 Is the Exchange Trustworthy?

The trustworthiness of the Exchange is one of the assumptions that the thesis makes. The premise is that the Exchange's primary function is to provide blind *proof-of-existence* to enable

Violence policy, and is no longer searchable on Twitter.

[3]https://www.trulioo.com

[4]https://kantarainitiative.org

[5]https://www.veriff.com

[6]https://www.onelogin.com

[7]https://www.au10tix.com

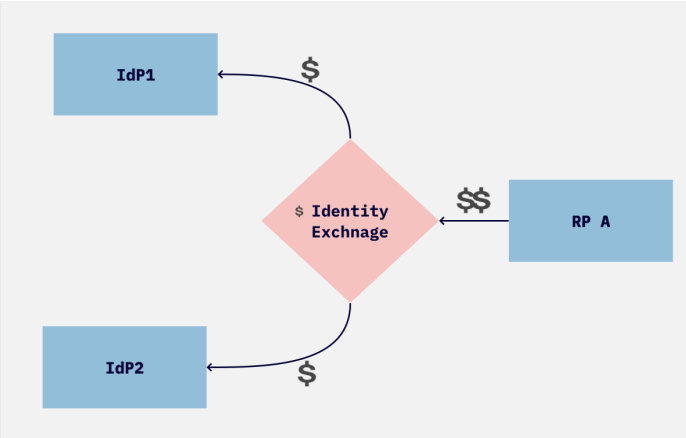

Figure 6-2: Brokered Federation's Business Proposition

anonymity-preserving accountability measures that are independent of the platform themselves. To maintain this service, they must provide accurate information, must follow all privacy guidelines, store minimal data and always ask for user consent. Their business proposition is the motivating factor to trust that the Exchange will not behave badly or else it will lose its customers (the RPs). This thesis does not do an in-depth analysis of the monetary incentives and divisions, however Figure 6-2 shows a bare bones idea of how the financial structure of the Identity Federation might come about. The RP will pay the Exchange money in accordance to a prior contract or based no the number of users of the RP that use the Exchange in the identity proofing procedure. Similarly, the Exchange will pay IdPs for services rendered based on the number of people that choose a particular IdP for identity proofing purposes. The RP will be motivated to pay the Exchange as curbing disinformation and adding cost to fake account creation is a security measure that might be worth investing in.

Chapter 7

Conclusion

Efforts to curb the growing threat of disinformation and social media influence campaigns via social bots are ever growing. Current detection techniques for social botnets on large platforms consistently find and remove disinformation campaigns; the most recent campaign reported by Facebook was a disinformation campaign that targeted countries in North Africa and the Middle East and originated from France and Russia[1]. This is an example of one of many such attacks which aim to direct public opinion to nefarious agendas. However, despite these efforts fake accounts continue to abuse these online community platforms. This is largely because the cost of creating a new account is minimal for bad actors.

Chapter 2 discussed the different types, sophistication and purpose of social bots and the current techniques used to detect them, which is a mix of technology and human review. Other than the reactive Social bot detection techniques, chapter 2 also covered other techniques to limit fake account creation such as reputation schemes which are common in E-commerce websites and online advice communities to add value to an account and hence increase of cost of losing an account which has high reputation points.

A low-hanging fruit that many large online community platforms are moving towards to counter this threat is to ask users for more personally identifiable information. More and more of these platforms require users' phone numbers. In case of supposed 'suspicious behaviour', users are now being requested to provide government ID card photos, copies of birth certificates

[1]https://about.fb.com/news/2020/12/removing-coordinated-inauthentic-behavior-france-russia/

or credit card information. This move is due to the incorrect assumption that identifiability is the only mechanism for accountability. Chapter 3 makes a case of how anonymity and accountability are not mutually exclusive and it is possible to create anonymous *and* accountable identity schemes online.

7.1 Contributions

This thesis makes a claim that a *proof-of-existence* is sufficient to establish that an account-holder is a real person. A platform does not need to know actual personal details about a user to provide security measures. For example, a platform looking to establish whether a user is over the age of 18, does not need the actual birth date of a user - just a Boolean that answers "Are you over the age of 18?". This should be sufficient for a platform as long as the answer is coming from a trusted source. To maintain anonymity, this *proof-of-existence* should not reveal excess information about a user to both the receiver of this proof and the sender of this proof (chapter 4).

A realization of such a double-blind concept is offered by the *Brokered Federation Model*. The thesis discussed existing Brokered Identity Federations in existence today in chapter 4 - used mostly in federal or government Identification context. The thesis's main contribution has been to design a Brokered Identity Federation architecture inspired from existing work and re-purposed it for a one-time blind *proof-of-existence*. Chapter 5, details how such an architecture would come about. The distinguishing features being:

- Using Brokered Identity Federation to off-load the task of Identity Proofing to Identity Service Providers which are separate entities from the Relying Parties that are requesting identity information about prospective users. The double-blind capability is provided by the Exchange that prevents tracking and profiling of users by the Relying Party and/or the Identity Service Provider.
- Providing a means to have stringent Identity Proofing (on request of Relying Party) without leaking any unnecessary information to the Relying Party. The Identity Proofing step ensures that there is a burden of proof for every new account, hence bad actors cannot use Social bots to freely create multitudes of accounts (large scale platform abuse).
- The Exchange keeps track of the number of accounts created on a Relying Party by the

same Identity Service Provider user, hence also making accounts less discardable for bad actors who are not using social bots (small scale platform abuse). After a certain limit, a Relying Party will not allow a bad actor to create more accounts based on only the knowledge of the count the previous accounts on the platform. They do not know who the person is, which previous accounts the person had.

- Provides a worked example of key identity interactions through the Exchange with the OpenID Connect (OIDC) protocol (Appendix B).
- Introduces the concept of *Trust Framework* which ensures compliance from a non-technical but a contractual and policy perspective (chapter 4).

The architecture in itself cannot guarantee success though. Considerations like usability and accessibility are of utmost importance when it comes to adoption of any new global systems. While the thesis provides general guidelines for the usability of a brokered model, however, it does not provide decisions on who should be the Exchange or who should be included as IdPs for a particular Identity Federations. The decision of the Exchange is important because to the user, the Exchange is the barrier that separates their sensitive information stored on the IdP from the *proof-of-existence* sent to the RP. Trust in the Exchange is paramount for the success of such a brokered Federation model. This thesis makes the assumption that the Exchange is trustworthy because their business proposition and the vital functionality is to preserve user anonymity. Moreover, the thesis does not specify who the IdPs should be but does discuss the important implications of choice of IdPs to ensure global accessibility in chapter 6, among other considerations that are vital to take into account.

The architecture is designed to provide a large margin of flexibility. There are knobs and switches that participants (especially the Relying Party) of the Federations can adjust to suit their requirements, in terms of IdP selection, Exchange selection, protocols usable, assurance levels for different types of identity transactions and the information shared among participants.

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix A

Appendix: Architecture Features

A.1 Terminology Mapping

Figure A-1 shows the terminology mappings across different literature on brokered identity federation.

Proposed Architecture Term	Australia's TDIF Term	NIST Term	UK's GOV.UK Verify Term	OIDC Term	SAML Term
Relying Party (RP)	Relying Party (RP)	Relying Party (RP)	Relying Party (RP)	Relying Party (RP)	Service Provider (SP)
Identity Service Provider(IdP)	Identity Service Provider(IdP)	Identity Provider (IdP)	Identity Provider (IdP)	OpenID Provider (OP)	Identity Service Provider(IdP)
	Credential Service Provider (CSP)	Credential Service Provider (CSP)			
	Attribute Provider (AP)	Identity Provider (IdP)			
Identity Exchange (IX)	Identity Exchange (IDX)	Proxied Federation	The Hub		

Figure A-1: Terminology comparison amongst different architecture documents

A.2 Levels of Assurances Explained

A.2.1 Identity Assurance Level (IAL)

Identity Assurance Levels are used in the context of Identity Proofing, a process to validate the correctness of the identity attributes of a natural person and verify that the attributes belong to the natural subject. While NIST[GGF17] has 3 IALs, Australia's TDIF[Age20] has 4 of these "Identity Proofing Levels" and UK's GOV.UK Verify[Whi18] has similarly 4 "Levels of Assurance - ID". To be on the more wary side, this proposal includes 4 levels of assurance to ensure that participating entities have the more flexibility to choose the assurance strength that works for them.

- **IAL1:** Identity attributes are self-asserted. Such attributes are not validated or verified.
- **IAL2:** In IAL2 there is evidence that identity attribute exists in the real world, however the association with the person presenting for validation is not the strongest.
- **IAL3:** Similar to IAL2, there is evidence of existence of identity attributes, but now there is a much stronger binding to the user presenting the evidence for validation.

- **IAL4:** In-person interview is required for identity proofing. The evidence of identity attributes is cross-verified against additional identity sources.

It is important to note that these IALs are for use by the user to prove their Identity to the IdP and none of the information provided to IdP will be shared with any RP without the explicit consent of the user. The purpose of having multiple IdPs with varying levels of trust is to ensure that the burden of any PII revealed to the IdP is proportional to the risk factor and necessity of the information needed rather than an overreaching action. Moreover, attributes asserted by the IdPs to the RPs can be used to support the pseudonymous identity on the RP, and not reveal any information held by the IdP itself.

A.2.2 Authenticator Assurance Level (AAL)

While IAL describes strength of identity proofing, or alternatively establishes that a person is who they claim to be, Authentication Assurance Level (AAL) describes the strength of authentication, or alternatively establishes that the person attempting to access a digital service is the owner of the identity. The term AALs comes from the NIST documentations[GGF17] but similar levels called "Authentication Credential Level" also exist in the Australian TDIF[Age20].

- **AAL1:** AAL1 requires a single factor of authentication, primarily username and password but not limited to them. It provides some level of confidence that the person controls the credential bound to their identity.
- **AAL2**: AAL2 requires at least two-factor authentication. Hence, provides higher assurance that the person authenticating is the owner of the user's identity on a digital service.
- **AAL3:** AAL3 provides very high assurance that the person authenticating is the owner of the user's identity on a digital service. It also requires further constraints in what credentials can be used and how to possess two different authentication factors (i.e something you know, something you have, and something you are).

In most if not all federated identity scenarios, the user does not authenticate directly to RP. Instead the credentials associated with the IdP are used to generate an assertion for an identifier associated with the user, defined by the federation framework, to gain access to the RP. However, that is not the case for the current proposal. The aim of the proposal is not to use IdP credentials to authenticate to a RP but use assertions provided by IdP to provide

information about a new user. The assertion provided by an IdP will be used as the last step for a user to sign up to an RP - where the assertion will provide proof of 'natural person ' attested by the IdP. Hence, the user only has to use the IdP once - when making a new account on an RP. For all subsequent authentications into the RP, it maintains separate credentials and hence does not need to go through the federation architecture.

A.2.3 Federation Assurance Level (FAL)

The term Federation Assurance Level (FAL) only exists in NIST 800-63-3[GGF17], to the best of my knowledge. FAL describes aspects of the assertion and federation protocol used in a given transaction.

- **FAL1:** Allows for the user to enable the RP to receive an assertion. The assertion is signed by the IdP using approved cryptography[GGF17].
- **FAL2:** Adds the requirement that the assertion be encrypted using approved cryptography such that the RP is the only party that can decrypt it[GGF17].
- **FAL3:** Requires the user to present proof of possession of a cryptographic key referenced in the assertion in addition to the assertion artifact itself. The assertion is signed by the IdP and encrypted to the RP using approved cryptography[GGF17].

At all FAL, the IdPs ensures that an RP (or any other malicious party) cannot impersonate the IdP at another RP by signing the assertion. This provides protection that no other entity than the private key holder (IdP) can sign the assertion. The IdP must publish its public key in a verifiable fashion, such as at an HTTPS-protected URL at a well-known location.

Based on IALs, AALs and FALs, RPs can decide which IdPs can be used to provide identity assertion for new users on their platform and vice versa IdPs can decide the RPs it is willing to accept requests from. This information can be pre-determined and updated in a periodic time to establish whitelists and blacklists for the participants in the Federation.

A.3 Threats and Mitigation

The threats, their descriptions as well as recommended mitigation strategies are reported from NIST 800-63-3[GGF17] in table A.1 .

Federation Threats	**Description**	**Mitigation Strategies**
proof-of-existence modified	Attacker creates or modifies an assertion	IdP should sign the *proof-of-existence* cryptographically. Salt the assertion. Add a non-guessable identifier.
proof-of-existence re-directed	Assertion passed to attacker instead of intended party by the user agent.	Include identity of the receiver as the name of the Exchange or the RP. Some protocols remove the role of user agent in identity transactions (eg: OIDC in authorization code flow configuration)
proof-of-existence re-used	Assertion from Exchange to RP re-used by attacker for their own session	Add timestamp and minimized validity period.
proof-of-existence substituted	Session hijacking attack	Some protocols remove the role of user agent in identity transactions (eg: OIDC in authorization code flow configuration), hence removing possibility of hijacking
proof-of-existence leaked	Attacker is able to intercept assertion and view it	Encrypt the assertion.
proof-of-existence repudiated by IdP	IdP refuses that they did any transaction	IdP should sign the *proof-of-existence* cryptographically with a key that allows for non-repudiation.

Table A.1: Threats and Mitigation strategies

THIS PAGE INTENTIONALLY LEFT BLANK

Appendix B

Appendix: OpenID Connect (OIDC)

B.1 OIDC and how it works

On the most rudimentary level, OIDC is a security mechanism which can be used by a Relying Party to request identity information from an Identity Provider. It's main feature is an authorization mechanism which allows one party to access and use information from another party.

OIDC, like Oauth and SAML, has design features which ensures that a bad actor can't steal information by pretending to be someone else such as specifying the receiver as `client_id`, the receiver endpoint as `redirect_uri` and having a random secret as `state` for a particular session. The `scope` or the information to be shared is also pre-determined between the parties.The specification suite for OIDC is extensive and details can be found on their official website[1].

For the purpose of this document, only information relevant to the proposed architecture design will be described in depth while the different variations possible within the OIDC protocol suite will be ignored. Moreover, details of underpinning elements such as OAuth, OAuth 2.0 and JWT will also be ignored. The assumption is that the reader must trust these elements just as the reader trusts encryption or TLS.

[1] https://openid.net/connect/

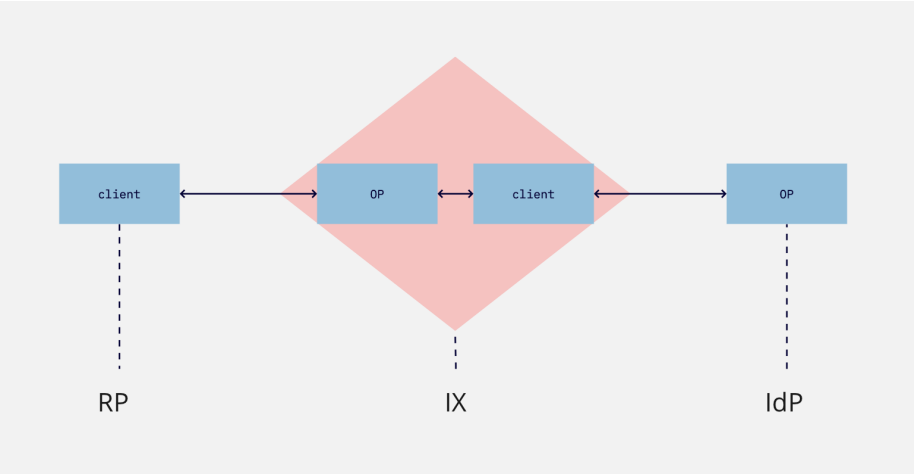

Figure B-1: Brokered Federation Model and OIDC mapping

Most people have probably interacted with a system that uses OIDC in the background. For example, a user accessing The New Yorker Magazine through their Facebook account has OIDC in the background. In this scenario Facebook acts as the *OpenID Provider* (OP) and the service being accessed is the Relying Party or the client, which is The New Yorker Magazine.

I purposely dissociate between the OIDC terms (client and OpenID Provider) from the brokered federation model terms (IdP, IX, and RP) because the IX plays the both the roles of client and OpenID Provider (OP) depending on whether the IX is interacting with the IdP or the RP. When the IX interacts with the RP it acts as the OpenID Provider and when the IX is interacts with the IdP it acts as the client. So, one successful interaction that passes between RP, IX and IdP requires two implementations of OIDC protocol. One that is between the RP and IX, and one that is between IX and IdP as highlighted in figure B-1.

B.2 OIDC based Worked Example

This example is spread across the figures B-2, B-3 and B-4 and a step-by-step guidance is provided in the details below.

1. The user discovers a relying party.

    1.1. The user access the RP with the intent to create an account and access services on

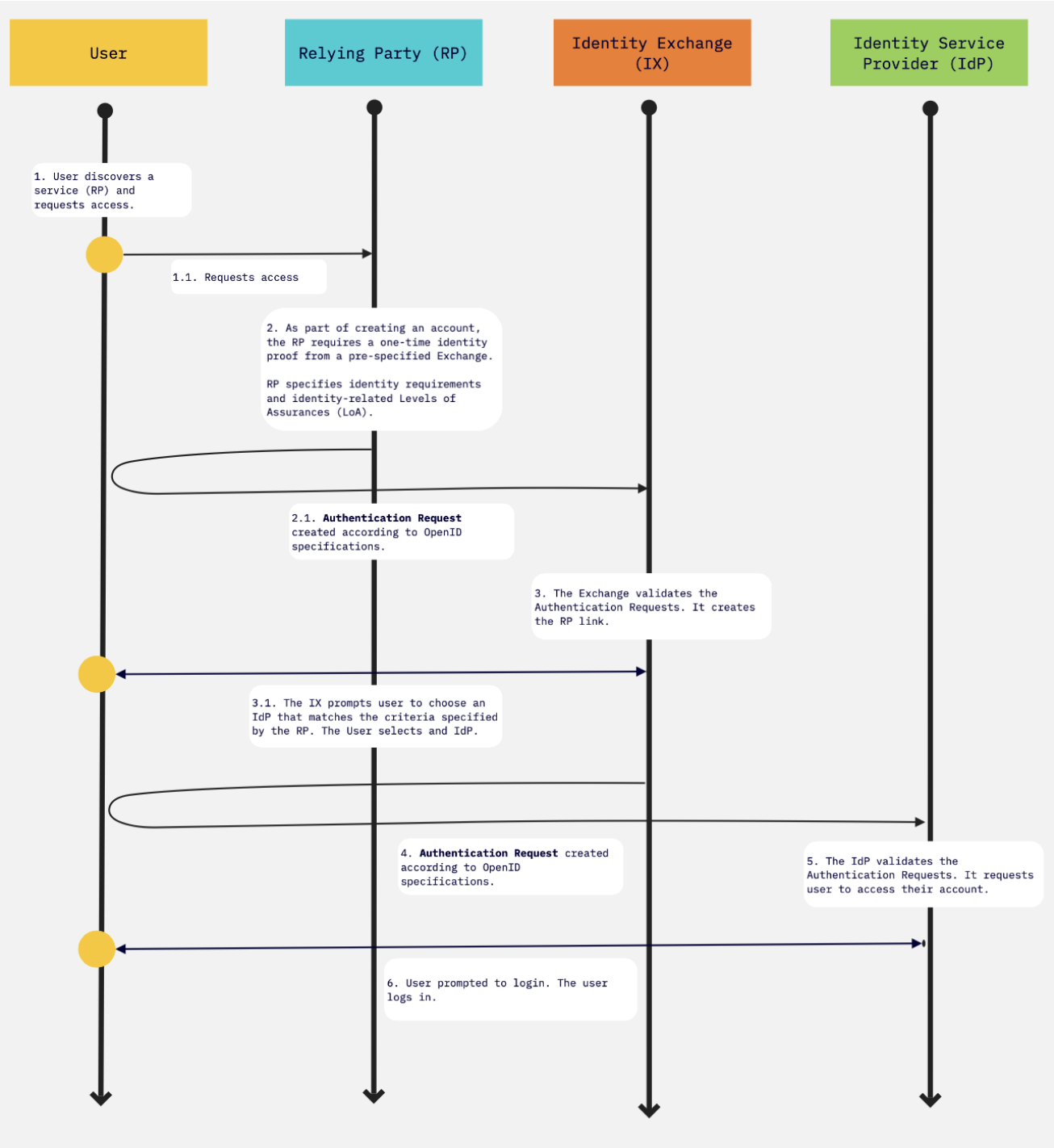

Figure B-2: Sequence Diagram (step 1 to 6)

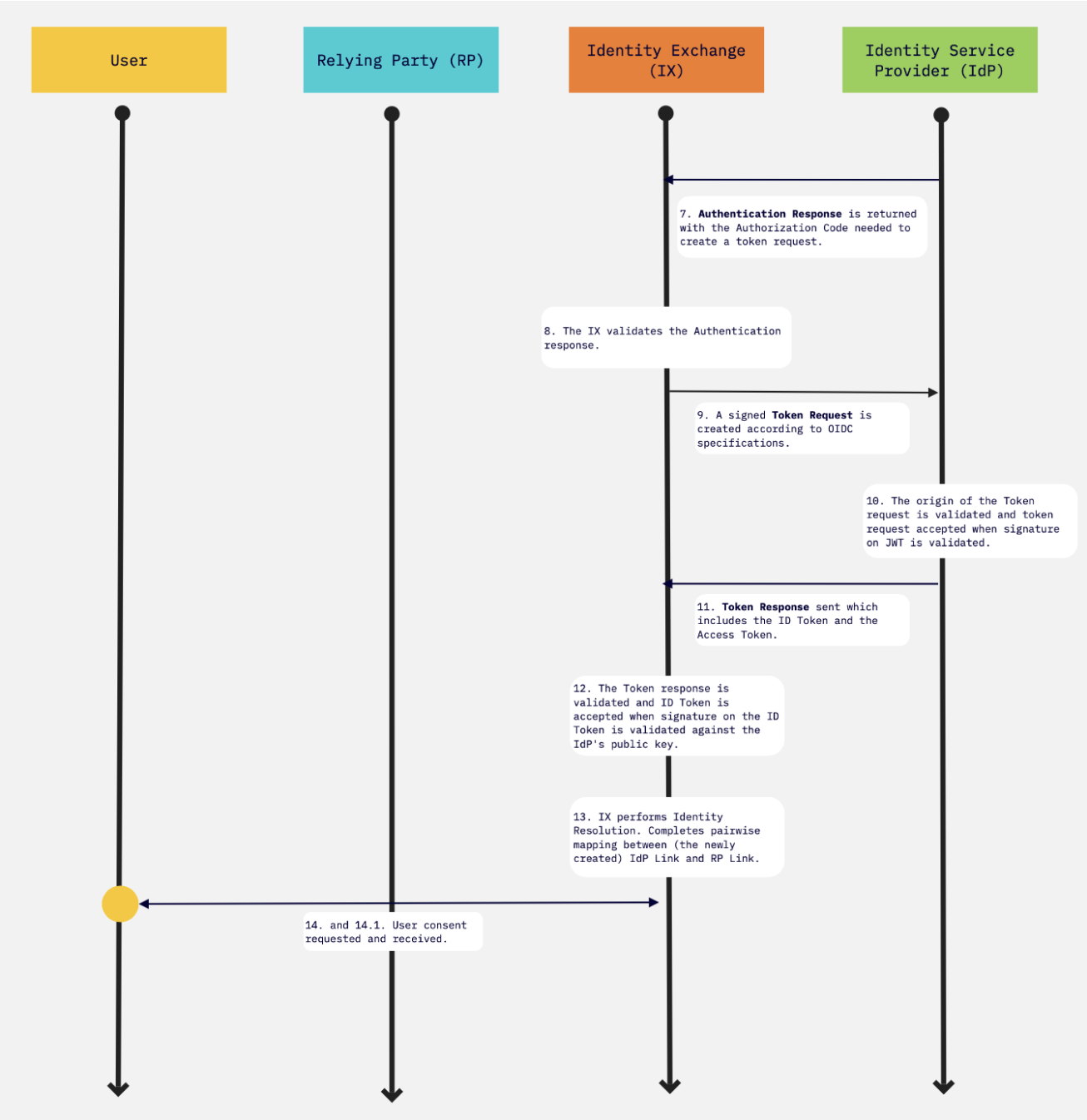

Figure B-3: Sequence Diagram (step 7 to 14)

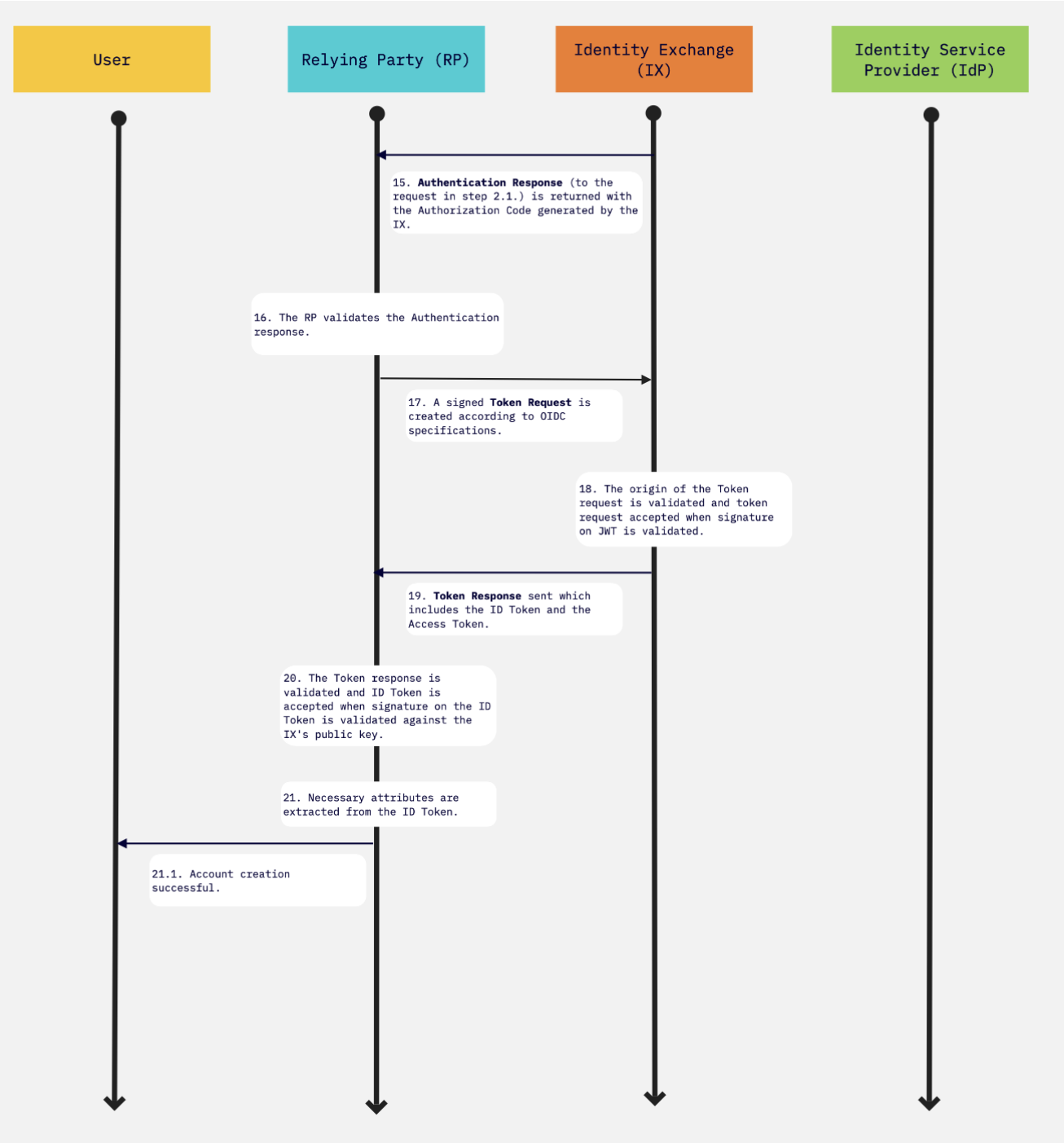

Figure B-4: Sequence Diagram (step 15 to 21)

the RP.

2. The user creates an account on the RP and as the last step of account creation, the RP asks for access to identity information from a user through a trusted IX.

   2.1. The Relying Party acting as the client, creates an authentication request including the scope parameters which includes any attributes requested, proof of attributes requested and the Level of Assurance (LoA) requirements. The authentication request includes:

   - A redirect response created by the client, which triggers the user agent to make an authentication request to the OpenID Provider (OP) which is the Exchange in this context.
   - This authentication request includes a `client_id` which is the client identifier, the `scope`, the `redirect_uri` where the client will receive the authentication response, the `state` which is a random string generated by the client to identify a session, prevent CSRF attacks and must be returned to the client in teh authentication response, and other optional specifications.

3. The IX logs the request from the RP and stores it against an identifier that it creates called the *RP Link*. The IX also validates the authentication request from the RP.

   - This RP Link will be used to connect the information requested by the RP to the information provided by the IdP without either party knowing the identity of each other.

   3.1. The IX prompts the user to select an IdP. The user selects an IdP.

4. Based on the selected IdP, the IX creates a Authentication Request for the selected IdP. *Now the Exchange acts as a client and the IdP will be the OP*. The IX will use all the information provided by the RP to create the `scope`. The authentication request includes:

   - A redirect response created by the client (the IX now), which triggers the user agent to make an authentication request to the OpenID Provider (OP) which is the IdP now.
   - This authentication request includes a `client_id` which is the client identifier (identifies the IX), the `scope`, the `redirect_uri` where the client will receive

the authentication response, the `state` which is a random string generated by the client to identify a session, prevent CSRF attacks and must be returned to the client in the authentication response, and other optional specifications.

5. The IdP validates that the authentication request came from the IX.
6. The IdP prompts user to log into their account. The user provides necessary credentials to access their account on the IdP.

    - There might be extra steps required for the user to satisfy the LoA requirements. These requirements will be specified in the `scope` of the authentication request. The Identity Assurance Level (IAL) or the Authenticator Assurance Level (AAL) might not be satisfied. The user will be required to meet the requirement levels by identity proofing mechanisms or adding multi-factor authentication to their IdP.

7. An authentication response is returned to the client (the IX) which includes a `code` generated by the OP (the IdP) and the same `state` value provided by the client in the Authentication request in step 4.

    - The authorization code called `code` is a random string issued by the IdP to be used in the request to the *token endpoint* - this is an endpoint at the OP (which in this step is the IdP). The OIDC has many code flows, for the purpose of this architecture I am using the Authorization code flow. The authorization code flow ensures that none of the tokens are exposed to the User Agent which removes the chance of any malicious applications on the User Agent being able to access the tokens.

8. The IX validates the authentication response.
9. Now, the IX creates a Token request which includes the `code` that was received in the authentication response in step 7, the `redirect_uri` which must match the value used in the authentication request in step 4, and the `client_assertion` which is the the signed client authentication JWT generated by the client(IX). The client must generate a new assertion JWT for each call to the token endpoint at the IdP.

    - The signed JWT is a Json Web Token which has claims made by the IX and signed by it as well. These claims include `iss` which is the client ID of the client creating and issuing the JWT, the `aud` which is the URL of the OP's (which is the IdP in this step) token endpoint, the `jti` which is a unique random identifier of the JWT and the date of creation and expiration.

10. The IdP validates the Token request. The Token request is accepted when signature on the JWT (`client_assertion`) is validated using the IX's registered public key.
11. The OP (IdP) returns a Token Response which includes an ID Token and an Access Token, signed by the IdP.

    - **ID Token**: It is the signed JWT which includes set of claims sent by the OP. These claims includes the `iss` which is the URL of the OP creating and issuing the ID Token (which is the IdP in this step, the `aud` which is client ID of the client (IX), the `sub` which is the identifier of the user, the `acr` which is the level of assurance at which the user was authenticated at, at the IdP, the `jti` which is a unique random identifier of the JWT to prevent reuse of token and the date of creation and expiration of the ID Token.
    - The `sub` is a pairwise unique value which identifies the end-user of the OP to the client only. So a different Exchange will have a different value for the `sub` identifier for the same end-user. This is added to remove linkability if two clients collude (which in this case is two Exchanges).
    - The ID Token may also include other requested claims (attributes and proof of attributes) such as the end user's email address or a Boolean confirming whether the end-user is over the age of 18.
    - **Access Token**: An Access token can be used to make further requests for more User Information. For the sake of this example, I assume the RP requires no excess identity attributes from the IdP.
12. The IX validates the Token response. The IX validates the ID Token and accepts it if the signature on the ID Token is validated using IdP's registered public key.
13. The IX extracts the subject identifier(`sub`) from the ID Token and checks whether it already has a an entry for that particular subject identifier. If one doesn't exist, The Identity Exchange creates one and stores it against an identifier generated by the Exchange called the *IdP link*.

    - Just like the RP Link mentioned in step 3, the IdP Link will be used to connect information provided by an IdP to an RP.
14. The IX extracts all other claims from the ID Token as well. Before passing on the attributes to the RP that requested them, the IX asks for user consent.

14.1. The IX prompts the user to give consent. The user provides the necessary consent.

15. The Authentication Response to the request created in step 2.1 is sent back to the client. *Now the Exchange is back to acting as the OP and the RP will be the client*. The response includes a `code` generated by the OP (the IX) and the same `state` value provided by the client in the Authentication request in step 2.1.
16. The RP validates the Authentication response.
17. The RP creates a Token Request to the IX. This request includes the `code` , the `redirect_uri` used in the authentication request and the `client_assertion` which is the signed JWT generated by the client (RP).
18. The IX validates the Token request. It also validates the signature on the JWT against the RP's registered public key.
19. The IX returns a successful Token response. This includes an ID Token and an Access Token signed by the IX. The ID Token includes all the information required by the RP in its initial request. The ID Token also contains the number of times this IdP has been used by the the end-user to create an account on the RP.
20. The RP validates the Token response, as well as the ID Token and the JWT signed in the ID Token.
21. The RP extracts all other claims from the ID Token as well. This includes identity attributes, proof of identity attributes, count of previous accounts on this RP and Levels of Assurance.

    21.1. The RP lets user know that account creation was successful if the previous count of accounts is less than the threshold created by the RP.

THIS PAGE INTENTIONALLY LEFT BLANK

Bibliography

[Age20] Digital Transformation Agency. *The Trusted Digital Identity Framework*. Oct. 2020. Retrieved from https://www.dta.gov.au/our-projects/digital-identity/trusted-digital-identity-framework.

[Aki11] Akky Akimoto. *Japan, the Twitter nation*. May 2011. Retrieved from https://www.japantimes.co.jp/life/2011/05/18/digital/japan-the-twitter-nation/.

[AlR+15] M. AlRubaian et al. "A novel prevention mechanism for Sybil attack in online social network". In: *2015 2nd World Symposium on Web Applications and Networking (WSWAN)*. Mar. 2015, pp. 1–6. DOI: 10.1109/WSWAN.2015.7210347.

[Arn18] Dan Arnaudo. *Brazil: Political Bot Intervention During Pivotal Events*. Oxford University Press, 2018.

[Atw09] Jeff Atwood. *A Day in the Penalty Box*. Apr. 2009. Retrieved from https://stackoverflow.blog/2009/04/06/a-day-in-the-penalty-box/.

[AYM15] Norah Abokhodair, Daisy Yoo, and David W. McDonald. "Dissecting a Social Botnet: Growth, Content and Influence in Twitter". In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW '15. Association for Computing Machinery, Feb. 2015, pp. 839–851. DOI: 10.1145/2675133.2675208.

[Baa+15] Tim Baarslag et al. "Negotiating mobile app permissions". In: May 2015.

[BF16] Alessandro Bessi and Emilio Ferrara. *Social Bots Distort the 2016 US Presidential Election Online Discussion*. ID 2982233. Nov. 2016.

[BH] Samantha Bradshaw and Philip N. Howard. *Computational Propaganda | The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*. Retrieved from https://comprop.oii.ox.ac.uk/research/

posts/the-global-disinformation-order-2019-global-inventory-of-organised-social-media-manipulation/.

[BM06] Jennifer Brown and John Morgan. "Reputation in Online Auctions: The Market for Trust:" in: *California Management Review* (Oct. 2006). DOI: 10.2307/41166371.

[Boe08] Tom Boellstorff. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human*. REV-Revised. Princeton University Press, 2008. DOI: 10.2307/j.ctvc77h1s.

[Bus20] Joline Buscemi. *Who's still on 'Second Life' in 2020?* Feb. 2020. Retrieved from https://www.mic.com/p/second-life-still-has-dedicated-users-in-2020-heres-what-keeps-them-sticking-around-18693758.

[Cac16] C. Cachin. *Architecture of the Hyperledger Blockchain Fabric*. 2016. Retrieved from https://www.semanticscholar.org/paper/Architecture-of-the-Hyperledger-Blockchain-Fabric-Cachin/f852c5f3fe649f8a17ded391df0796677a59927f.

[Cao+12] Qiang Cao et al. "Aiding the Detection of Fake Accounts in Large Scale Social Online Services". In: 2012, pp. 197–210.

[CG19] the Comptroller and Auditor General. *Investigation into Verify - National Audit Office (NAO) Report*. Mar. 2019. Retrieved from https://www.nao.org.uk/report/investigation-into-verify/.

[Che19] Brian X. Chen. "I Shared My Phone Number. I Learned I Shouldn't Have. (Published 2019)". In: *The New York Times* (Aug. 2019).

[Cre+19] Stefano Cresci et al. "Cashtag Piggybacking: Uncovering Spam and Bot Activity in Stock Microblogs on Twitter". In: *ACM Transactions on the Web* 13.2 (Apr. 2019), 11:1–11:27. DOI: 10.1145/3313184.

[Dav+16] Clayton A. Davis et al. "BotOrNot: A System to Evaluate Social Bots". In: *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion* (2016). arXiv: 1602.00975, pp. 273–274. DOI: 10.1145/2872518.2889302.

[DC20] Australian Data and Digital Council. *State of the Data and Digital Nation*. Sept. 2020. Retrieved from https://www.pmc.gov.au/resource-centre/public-data/state-data-and-digital-nation-september-2020.

[DiR+19] Renee DiResta et al. "The Tactics and Tropes of the Internet Research Agency". In: *U.S. Senate Documents* (Oct. 2019).

[Ely+13] Aviad Elyashar et al. “Homing socialbots: intrusion on a specific organization’s employee using Socialbots”. In: *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’13. Association for Computing Machinery, Aug. 2013, pp. 1358–1365. DOI: 10.1145/2492517.2500225.

[Ely+16] Aviad Elyashar et al. “Guided socialbots: Infiltrating the social networks of specific organizations’ employees”. In: *AI Communications* 29.1 (Jan. 2016), pp. 87–106. DOI: 10.3233/AIC-140650.

[Fei+11] Joan Feigenbaum et al. “Accountability and deterrence in online life”. In: *Proceedings of the 3rd International Web Science Conference*. WebSci ’11. Association for Computing Machinery, June 2011, pp. 1–7. DOI: 10.1145/2527031.2527043.

[Fer+16] Emilio Ferrara et al. “The rise of social bots”. In: *Communications of the ACM* 59.7 (June 2016), pp. 96–104. DOI: 10.1145/2818717.

[FJW11] Joan Feigenbaum, Aaron D. Jaggard, and Rebecca N. Wright. “Towards a formal model of accountability”. In: *Proceedings of the 2011 New Security Paradigms Workshop*. NSPW ’11. Association for Computing Machinery, Sept. 2011, pp. 45–56. DOI: 10.1145/2073276.2073282.

[GAA18] Christian Grimme, Dennis Assenmacher, and Lena Adam. “Changing Perspectives: Is It Sufficient to Detect Social Bots?” In: *Social Computing and Social Media. User Experience and Behavior*. Ed. by Gabriele Meiselwitz. Lecture Notes in Computer Science. Springer International Publishing, 2018, pp. 445–461. DOI: 10.1007/978-3-319-91521-0_32.

[GGF17] Paul A. Grassi, Michael E Garcia, and James L Fenton. *NIST Special Publication 800-63-3*. June 2017. Retrieved from /sp800-63-3.html.

[Gli19] Bryan Glick. *Why Gov.uk Verify faces a critical few months - again - Computer Weekly Editors Blog*. Aug. 2019. Retrieved from https://www.computerweekly.com/blog/Computer-Weekly-Editors-Blog/Why-Govuk-Verify-faces-a-critical-few-months-again.

[GVG15] Oana Goga, Giridhari Venkatadri, and Krishna P. Gummadi. “The Doppelganger Bot Attack: Exploring Identity Impersonation in Online Social Networks”. In: *Proceedings of the 2015 Internet Measurement Conference*. IMC ’15. Association for Computing Machinery, Oct. 2015, pp. 141–153. DOI: 10.1145/2815675.2815699.

[HB15] Eva Galperin Hassine and Wafa Ben. *Changes to Facebook's "Real Names" Policy Still Don't Fix the Problem*. Dec. 2015. Retrieved from https://www.eff.org/deeplinks/2015/12/changes-facebooks-real-names-policy-still-dont-fix-problem.

[Hei] Grant Heinich. *Bots*. Retrieved from https://bbcnewslabs.co.uk/projects/bots/.

[How+19] Philip Howard et al. "The IRA, Social Media and Political Polarization in the United States, 2012-2018". In: *U.S. Senate Documents* (Oct. 2019).

[Jee16] Charlotte Jee. *UK government identity scheme GOV.UK Verify launched today: What is GOV.UK Verify? GOV.UK Verify explained*. May 2016. Retrieved from https://www.computerworld.com/article/3426967/uk-government-identity-scheme-gov-uk-verify-launched-today--what-is-gov-uk-verify--gov-uk-verify-exp.html.

[Kan07] Colleen Kane. *The latest controversies surrounding Yelp*. Apr. 2007. Retrieved from https://fortune.com/2015/04/07/yelp-reviews-controversy/.

[Kar20] David Kariuki. *Pandemic spurs Second Life usage, book club, lower non-profit prices – Hypergrid Business*. Apr. 2020. Retrieved from https://www.hypergridbusiness.com/2020/04/second-life-sees-increase-in-users-during-coronavirus-pandemic/.

[Kir19] Sarah Kirk-Douglas. *SecureKey and Initial Network Participants Accomplish Key Milestone in Bringing Digital Identity Network to Market*. May 2019. Retrieved from https://verified.me/verifiedme-launch-release/.

[Lam05] Butler Lampson. *Accountability and Freedom Slides*. Sept. 2005. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.187.8982.

[LMR13] Ewa Luger, Stuart Moran, and Tom Rodden. "Consent for all: revealing the hidden complexity of terms and conditions". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. Association for Computing Machinery, Apr. 2013, pp. 2687–2696. DOI: 10.1145/2470654.2481371.

[LZ16] Michael Luca and Georgios Zervas. "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud". In: *Management Science* 62.12 (Dec. 2016), pp. 3412–3427. DOI: 10.1287/mnsc.2015.2304.

[Mar99] Gary T. Marx. "What's in a Name? Some Reflections on the Sociology of Anonymity". In: *The Information Society* 15.2 (May 1999), pp. 99–112. DOI: 10.1080/019722499128565.

[Mic19] Casey Michel. "Opinion | How the Russians pretended to be Texans — and Texans believed them". In: *Washington Post* (2019).

[MM19] Steven Lee Myers and Paul Mozur. "China Is Waging a Disinformation War Against Hong Kong Protesters (Published 2019)". In: *The New York Times* (Aug. 2019).

[MST17] Esther Makaay, Tom Smedinghoff, and Don Thibeau. *Trust Frameworks for Identity Systems - General area - OIX*. June 2017. Retrieved from https://openidentityexchange.org/networks/87/item.html?id=175.

[Mul00] Richard Mulgan. "'Accountability': An Ever-Expanding Concept?" In: *Public Administration* 78.3 (2000), pp. 555–573. DOI: https://doi.org/10.1111/1467-9299.00218.

[Mur+16] Dhiraj Murthy et al. "Automation, Algorithms, and Politics| Bots and Political Influence: A Sociotechnical Investigation of Social Network Capital". In: *International Journal of Communication* 10.0 (Oct. 2016), p. 20.

[Ora+20] Mariam Orabi et al. "Detection of Bots in Social Media: A Systematic Review". In: *Information Processing and Management* 57.4 (July 2020), p. 102250. DOI: 10.1016/j.ipm.2020.102250.

[Paw20] Krista Pawley. *Newly Launched Digital ID Framework to Begin Testing in Canada*. Sept. 2020. Retrieved from https://www.businesswire.com/news/home/20200915005744/en/Newly-Launched-Digital-ID-Framework-to-Begin-Testing-in-Canada.

[PK01] Andreas Pfitzmann and Marit Köhntopp. "Anonymity, unobservability, and pseudeonymity; a proposal for terminology". In: *International workshop on Designing privacy enhancing technologies: design issues in anonymity and unobservability*. Springer-Verlag, Jan. 2001, pp. 1–9.

[Rat+11] J. Ratkiewicz et al. "Detecting and tracking political abuse in social media". In: *In Proceedings of the 5th AAAI International Conference on Weblogs and Social Media (ICWSM'11*. 2011.

[Res+06] Paul Resnick et al. "The value of reputation on eBay: A controlled experiment". In: *Experimental Economics* 9.2 (June 2006), pp. 79–101. DOI: 10.1007/s10683-006-4309-2.

[SAV14] S. Sebastian, S. Ayyappan, and P. Vinod. "Framework for design of Graybot in social network". In: *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Sept. 2014, pp. 2331–2336. DOI: 10.1109/ICACCI.2014.6968575.

[SCM] Tao Stein, Roger Chen, and Karan Mangla. *Facebook Immune System*. Retrieved from https://research.fb.com/publications/facebook-immune-system/.

[Ser] Government Digital Services. Retrieved from https://www.gov.uk/government/publications/introducing-govuk-verify/introducing-govuk-verify.

[Sov18] Sovrin. *Sovrin: A Protocol and Token for Self-Sovereign Identity and Decentralized Trust*. Jan. 2018. Retrieved from https://sovrin.org/library/sovrin-protocol-and-token-white-paper/.

[Ste19] R. Stengel. *Information Wars: How We Lost the Global Battle Against Disinformation and What We Can Do About It*. Grove Atlantic, 2019.

[Sub+16] V. S. Subrahmanian et al. "The DARPA Twitter Bot Challenge". In: *Computer* 49.6 (June 2016), pp. 38–46. DOI: 10.1109/MC.2016.183.

[Sul19] Mark Sullivan. *Twitter admits it used info shared for security to target ads*. Oct. 2019. Retrieved from https://www.fastcompany.com/90415169/twitter-admits-it-used-info-shared-for-security-to-target-ads.

[Sun+12] San-Tsai Sun et al. *A Investigating User ' s Perspective of Web Single Sign-OnFIX ME!!!!: Conceptual Gaps , Alternative Design and Acceptance Model*. 2012. Retrieved from https://www.semanticscholar.org/paper/A-Investigating-User-%E2%80%99-s-Perspective-of-Web-Single-Sun-Pospisil/12bc9c3b54de80cbe7b4d3d6c831fabc1

[TCH19] LENE TCHEKMEDYIAN. *Man posed as teenager to have vulgar chats with girls on TikTok app, authorities say*. Feb. 2019. Retrieved from https://www.latimes.com/local/lanow/la-me-ln-tik-tok-lewd-acts-arrest-20190214-story.html.

[Tho12] Cadie Thompson. *Facebook: About 83 Million Accounts Are Fake*. Aug. 2012. Retrieved from https://www.cnbc.com/2012/08/02/facebook-about-83-million-accounts-are-fake.html.

[Utz+19] Christine Utz et al. "(Un)informed Consent: Studying GDPR Consent Notices in the Field". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. Association for Computing Machinery, Nov. 2019, pp. 973–990. DOI: 10.1145/3319535.3354212.

[Vaa15] Lisa Vaas. *Facebook finally changes real-name policy*. Nov. 2015. Retrieved from https://nakedsecurity.sophos.com/2015/11/03/facebook-finally-changes-real-name-policy/.

[Vap+15] Anna Vapen et al. "Information Sharing and User Privacy in the Third-Party Identity Management Landscape". In: *ICT Systems Security and Privacy Protection*. Ed. by Hannes Federrath and Dieter Gollmann. IFIP Advances in Information and Communication Technology. Springer International Publishing, 2015, pp. 174–188. DOI: 10.1007/978-3-319-18467-8_12.

[Var+17] Onur Varol et al. "Online Human-Bot Interactions: Detection, Estimation, and Characterization". In: *arXiv:1703.03107 [cs]* (Mar. 2017). arXiv: 1703.03107.

[Vos04] Marco Voss. "Privacy Preserving Online Reputation Systems". In: *Information Security Management, Education and Privacy*. Ed. by Yves Deswarte et al. IFIP International Federation for Information Processing. Springer US, 2004, pp. 249–264. DOI: 10.1007/1-4020-8145-6_20.

[VRA18] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (Mar. 2018), pp. 1146–1151. DOI: 10.1126/science.aap9559.

[Wan+12] Gang Wang et al. "Social Turing Tests: Crowdsourcing Sybil Detection". In: *arXiv:1205.3856 [physics]* (Dec. 2012). arXiv: 1205.3856.

[Wan10] Alex Hai Wang. "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach". In: *Data and Applications Security and Privacy XXIV*. Ed. by Sara Foresti and Sushil Jajodia. Lecture Notes in Computer Science. Springer, 2010, pp. 335–342. DOI: 10.1007/978-3-642-13739-6_25.

[War15] Rossalyn Warren. *Here's Why Having To Use "Real" Names On Facebook Is Putting Some People's Lives At Serious Risk*. June 2015. Retrieved from https://www.buzzfeed.com/rossalynwarren/heres-why-having-to-use-real-names-on-facebook-is-putting-so.

[Whi18] Edgar A Whitley. *Trusted Digital Identity Provision: GOV.UK Verify's Federated Approach*. Nov. 2018. Retrieved from https://www.cgdev.org/publication/trusted-digital-identity-provision-gov-uk-verify-federated-approach.

[Wil15] Lauran C Williams. *The Truth Behind Facebook's Real Name Policy*. July 2015. Retrieved from https://archive.thinkprogress.org/the-truth-behind-facebooks-real-name-policy-394196507b4f/.

[Wol12] Josephine Charlotte Paulina Wolff. “Unraveling Internet identitiesFIX ME!!!!: accountability and anonymity at the application layer”. Doctoral dissertation. Massachusetts Institute of Technology, 2012.