Manchester Metropolitan University

e-space

Manchester Metropolitan University's Research Repository

Please cite the Published Version

Yang, Yifan, Cooper, Daniel, Collomosse, John, Dragan, Constantin Catalin, Manulis, Mark, Steane, Jamie, Manohar, Arthi, Briggs, Jo, Jones, Helen and Moncur, Wendy (2022) TAPESTRY: a de-centralized service for trusted interaction online. IEEE Transactions on Services Computing, 15 (3). pp. 1385-1398. ISSN 1939-1374

DOI: https://doi.org/10.1109/tsc.2020.2993081

Publisher: Institute of Electrical and Electronics Engineers

Version: Accepted Version

Downloaded from: https://e-space.mmu.ac.uk/632090/

Additional Information: © 2020 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

TAPESTRY: A De-centralized Service for Trusted Interaction Online

Yifan Yang, Daniel Cooper, John Collomosse, Constantin C. Drăgan, Mark Manulis, Jamie Steane, Arthi Manohar, Jo Briggs, Helen Jones, Wendy Moncur

Abstract—We present a novel de-centralised service for proving the provenance of online digital identity, exposed as an assistive tool to help non-expert users make better decisions about whom to trust online. Our service harnesses the digital personhood (DP); the longitudinal and multi-modal signals created through users' lifelong digital interactions, as a basis for evidencing the provenance of identity. We describe how users may exchange trust evidence derived from their DP, in a granular and privacy-preserving manner, with other users in order to demonstrate coherence and longevity in their behaviour online. This is enabled through a novel secure infrastructure combining hybrid on- and off-chain storage combined with deep learning for DP analytics and visualization. We show how our tools enable users to make more effective decisions on whether to trust unknown third parties online, and also to spot behavioural deviations in their own social media footprints indicative of account hijacking.

Index Terms—Decentralised Trust, Online Identity, Artificial Intelligence, Interaction Design.

1 Introduction

Online fraud and scams are sharply on the increase, costing the global economy in excess of US$3 trillion in 2018 [60], and are often perpetrated through ephemeral false identities. Users struggle to make decisions on who to trust online, exposing themselves to risks from inappropriate over-disclosure of personal data. This motivates new techniques for determining the provenance and trustworthiness of digital identities – people, businesses or services – encountered online.

This paper reports on the outcomes of two years' work on the TAPESTRY project (EPSRC EP/N02799X/1), focusing upon a novel decentralised service that harnesses the complex longitudinal and multi-modal signals within citizens' digitally mediated interactions (for example, on social media) to support safe online interactions. The signals created via digital platforms – photos shared, comments left, posts 'liked' etc. – weave a complex 'tapestry' reflecting our relationships, personality and identity, referred to as the 'Digital Personhood' (DP). Commodification of the DP now fuels a billion-dollar industry in which machine learning is increasingly utilised to help make sense of, and extract value from, the deluge of DP data siloed for example within social platforms. In this work we exploit the DP for social good; through a platform (herein referred to as 'TAPESTRY' [1]) that empowers users to share 'trust evidence' of their DP in a granular, privacy preserving manner, in order to prove the provenance of their digital identity and so engender trust online. Our system enables a move away from a centralized, siloed model for personal data (and derived trust evidence) to a secure decentralised model that leverages distributed ledger technology (DLT), enabling users to retain agency over their data and to whom it is disclosed. Notably, TAPESTRY does not seek to make trust decisions on behalf of users. It is a decision support service, that enables granular exchange of trust evidence and its visualization in a summary form, to enable better human decisions to be made on trust.

We draw distinction between the problem of proving identity (authentication), and the problem tackled here, of proving the provenance of a digital identity. Online security is typically reliant on traditional representations of identity, taking simple pseudonyms or email addresses 'at face value' as users interact with one another or with digital services. We are now entering a new era in which citizens will construct a DP from childhood, comprising rich lifelong digital trails from social media and interactions with technology [36]. Those accumulated signals offer an increasingly viable way to prove the veracity of a digital identity. Leveraging the DP for this purpose poses significant challenges around signal processing, privacy, information security and infrastructure. Further challenges arise by designing the service for non-experts, who may have low levels of digital literacy - especially around numeracy. A fundamental tenet of the TAPESTRY platform is the preservation of the end-user as the owner of their trust decisions; we do not wish to develop a 'trust traffic light' or trust scoring system. Rather we wish to summarise in an intuitive way the trust evidence disclosed from one user to another, in order to support strong decision making around trust using that evidence. TAPESTRY tackles these challenges through three novel technical contributions:

1) A **secure data architecture** combining off-chain storage of encrypted trust evidence derived from the DP, backed by an unpermissioned proof-of-work (PoW) blockchain to ensure the integrity and

- *Y. Yang, D. Cooper and J. Collomosse are with the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey, UK.*
- *C. C. Drăgan and M. Manulis are with the Surrey Centre for Cyber-Security (SCCS), University of Surrey, UK.*
- *J. Steane and J. Briggs are with the School of Design, Northumbria University, UK. A. Manohar was with Northumbria University at the time this work was undertaken and is now at Brunel University, UK.*
- *W. Moncur is with the Duncan Jordanstone Centre for Art and Design (DJCAD), University of Dundee, UK. H. Jones was with DJCAD at the time this work was undertaken and is now at UCLan, UK.*

Corresponding Author: yifan.yang@surrey.ac.uk.

Manuscript received October 30, 2019.

1. Please note the distinction between our TAPESTRY project and platform and Zhao et al.'s Tapestry model for service deployment [64].

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

provenance of that evidence. The architecture incorporates a symmetric key sharing scheme, enabling granular disclosure of trust evidence. This provides users with agency over whom evidence is disclosed to, as well as control over the time periods and kinds of DP activity disclosed. (Secs. 3, 5).

2) A **machine learning** (ML) algorithm to irreversibly gist DP activity into compact descriptors that serve as the basic unit of trust evidence for sharing in the platform. We propose a deep neural network (DNN) to extract this evidence through a combination of semantic embedding and temporal modelling, enabling behavioural deviation to be detected over time. This in turn enables quantification of the regularity and temporal coherence of trust evidence which, combined with assurances over provenance and integrity from the blockchain, serves as the basis for users to make better trust decisions (Sec. 4). Although in principle trust evidence may be extracted from any social media modality, the scope of this paper focuses on textual posts from Twitter.
3) A **data visualisation** technique for representing the regularity and coherence of the trust evidence disclosed by a user within a single static image. The design of the visualisation is evaluated and shown to enable non-expert users to quickly make accurate determinations of the trustworthiness of a digital identity previously unknown to them (Sec. 6).

In order to evaluate our technical prototype of the TAPESTRY service, we explore two user-centric case studies where valid trust judgements and the avoidance of either fraud or victimisation are desirable for users.

First, we explore the efficacy of TAPESTRY to help users to detect fraudulent profiles in the context of crowdfunding within the video games industry. An early account of this evaluation experiment was given in a short workshop paper at IEEE Vizsec 2018 [62], but without any of the technical functionality operating behind the service, or discussion of the platform. Crowdfunding is a common vehicle by which small video games studios obtain financial support for new projects, and an online interaction in which investors must consider the trustworthiness of pitchers as a primary factor in making an investment [33]. We developed a controlled, workshop based evaluation of TAPESTRY in which the platform was used as an aid to investment decision-making within a mocked-up crowdfunding scenario. In this scenario, we used TAPESTRY to visualize trust evidence derived from real-world DPs of games developers, and artificial profiles fabricated and curated in the months prior to the study. We show that whilst TAPESTRY users do not make materially different trust decisions in terms of their accuracy (distinguishing the provenance of fake versus real identities), they are able to do so significantly more quickly using TAPESTRY, leading to advantages when making decisions online in time-pressured and information-overloaded situations.

Second, we explore the efficiency of TAPESTRY to help users detect unusual patterns of behaviour within their own DPs, pointing to unauthorized use (or account 'hijacking'). Again our goal is not to automatically raise an alarm or classify this behaviour, but to visually gist the trust evidence derived from a social media profile and the Digital Personhood that underpins it. When accustomed to the visual 'look and feel' of TAPESTRY visualisations of this trust evidence, we show that users can perceive deviations from the norm and so spot unusual patterns in online activities posted under their DP.

2 Related Work

Open authentication models (e.g. OAuth2) exist for establishing cross-site login without credential sharing, relying upon a trusted identity provider (e.g. Google, Facebook) to approve access to a digital identity. In addition, there is a range of services which help to establish trust, for both named and anonymous/pseudonymous users. For example, Escrow is a contractual arrangement used within the Dark Web, facilitated via a third party, which engenders trust between buyer and seller for crypto-currency transactions [44]. However, TAPESTRY is not proposing yet another access control solution or service for hosting digital identities or the DP, or for facilitating trusted transactions. Rather, TAPESTRY proposes an entirely new kind of service through which one may verify the trustworthiness of a digital identity through evidence derived from signals within an identity's DP.

2.1 Signals for Online Trust

The nature of trust is complex. It is '...a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behaviours of another' [16]. Importantly, this definition positions risk as naturally co-existing with trust: an individual accepts that the other party in an interaction may or may not act in the expected manner, but believes that their intentions are good. Offline, judgments about trust are informed by routinely available emotional and behavioural cues [10], [28], [54]. Online, these cues are usually absent. However, recent work indicates that there are alternate factors that may inform trust judgments [32], [33].

In the context of our crowdfunding case study of TAPESTRY (c.f. Sec. 7.2), prior work indicates that these alternate factors include (i) 'herding', where (e.g.) potential investors are reassured by the behaviour of previous investors, on the assumption that if others are doing something, it must be the rational thing to do [11], [39] and (ii) 'social proof', where (e.g.) less-expert investors are encouraged to invest later in a campaign by the involvement of early investors who are experts in product development or financial investment [35]. A further factor is social engagement. For example, trust is generated when creators of a crowdfunding campaign provide investors with updates on positive progress towards published goals [33]. This reassures investors and – indicative of trust – increases their investment [30], [39]. Trust is also generated when creators link their social media accounts [57]: investors likely feel that the creator has nothing to hide. Although such observations exist in the literature, including in an earlier account of our crowdfunding experiment as work-in-progress paper at IEEE Vizsec [62], TAPESTRY is unique in aggregating evidence from such sources to aid the user in their decision making on trust.

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

2.2 Social Media Analytics for Trust

Research increasingly explores opportunities for authentication of identity, making use of DP-related data on (e.g.) users' behaviours, activities, social media posts and search histories [37], [61]. Such data play a crucial role across many digital economy services including user profiling [21], personality [40] and crowdfunding [48]. Therefore, it is vitally important to protect DP by early detection of any malicious activities in social media feeds, to prevent economic or reputational harm. There exists various research methods in social media analytics for trust. Chalapathy et al. [9] and Yu et al. [63] explore detection of abnormal behaviours from regular group patterns, while Kang et al. [34] detect anomalous events through use of relationship graphs which model social network activities. In [1], the authors use social graph and text information to detect fraudulent comments in online review systems. Phua et al. [50] focus on structural metadata within posted social activities, instead of content. However, modeling users' behavioural norms in social media over longitudinal time periods, as well as visually representing this analysis to end-users, remains an open challenge that our research aims to address.

2.3 Trust and Identity over Blockchain

Blockchain's innovation is in its facilitation of direct transfer of unique digital property (e.g. currency, data, certificates) – previously reliant on third party intermediaries [2], [19]. This promotes 'trustless trust' [2] whereby exchanges are 'unidirectionally' trustworthy, and 'interpersonal trust' – trust in another online agent – is replaced through the technology's functionality of transparency, codification and immutability as a cryptographic audit trail [19]. However, while Blockchain brings significant trust-related functionality, many questions remain for end users about (i) how to demonstrate and prove such trust to an end user, and (ii) if – and if so how – a blockchain service enables this over existing intermediaries. Elsden et al. [19] catalogued over 200 blockchain applications and found that amongst identity management systems, most digital identities were provided by a third party (e.g Facebook or email account) or required supplementary state-backed documentation (passport, social security numbers etc.) to prove an identity. Amongst 'self-sovereign' digital identity Blockchain services, where a user issues and controls their own identity, many involved biometrics (e.g. fingerprint or iris scan) supported by other personal information (email address, bank details etc.). Dunphy and Petitcolas's in-depth review of identity management models using DLT [18] also found a prevalence of reliance on intermediaries, with the authors additionally summarising current UK and EU regulatory challenges i. e. 'know your customer'; anti-money laundering; and data protection. There are additional challenges of supporting the demonstration of the technology's unique trust-supporting benefits, and of communicating DP data in a visual form that supports intelligibility amongst non-expert users. TAPESTRY addresses this challenge through designing visualisations of trust evidence collected from social media, that communicate the coherence (and by extension, provenance) of the DP without disclosing specifics about a subject's past activities. Related to TAPESTRY's ability to evidence social media activity in a privacy preserving manner are distributed privacy-preserving social networks, such as Safebook [13] and early attempts to realize the functionalities of a social network using cryptographic techniques [4], [27]. The focus of these approaches has been on the privacy of users, and allowing them full control over their data. Thus, these approaches involve allowing the user to alter their data after previously committing to it. By contrast, TAPESTRY provides an immutable record of past social media activity, that a user may share to evidence the provenance of their digital identity. The use of Blockchain to provide such a service – analogous to a de-centralised credit reference for identity – is unique to TAPESTRY and addresses emerging concerns among the general public around the risks to privacy and security of siloing data within centralised services or organisations.

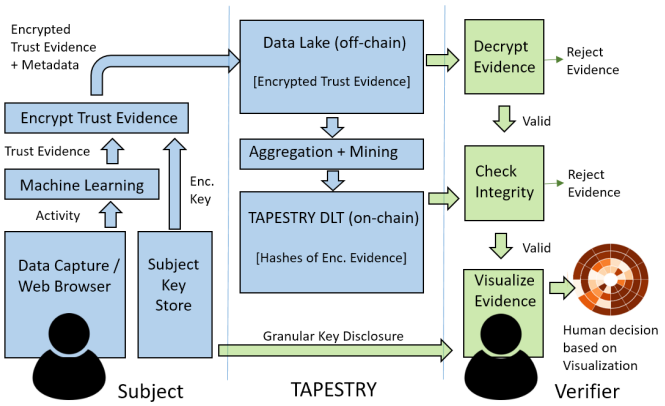

Fig. 1. TAPESTRY System architecture. Data is collected on subjects via the blue path. Digital activities (such as social media interactions) are captured on opt-in basis via a web browser extension. Trust evidence is derived from those activities via a deep neural network (DNN) and encrypted using a secret key. Keys differ between activities and change over time. Encrypted evidence is stored within a data lake, alongside metadata identifying the owner, timestamp and type of activity. A hash of that metadata serves as a unique ID to that evidence. The encrypted evidence is hashed alongside its unique ID within a proof-of-work blockchain. The green path enables a verifier to check the provenance of a subject by requesting disclosure of the relevant decryption keys. Encrypted evidence is requested from the lake; the provenance of that evidence is checked via the blockchain and it is decrypted. Analysis of the DNN signals yields a visualisation gisting the relevant period of activity that helps the user make a trust decision on the subject.

3 Overview of the TAPESTRY Service

The TAPESTRY service collects signals derived from the digital activities of a user (the 'subject') on an opt-in basis, and enables that subject to securely share those signals with another user (the 'verifier') in order to demonstrate the provenance of that subject's identity. We assume that these signals or 'trust evidence' (TE) are collected over longitudinal time periods from a rich tapestry of activities such as social media posts on various social platforms.

The verifier will determine what kinds of TE are sufficient to make a human decision on the trustworthiness of a user according to their use context.

The nature and quantity of trust evidence (TE) requested may vary considerably between contexts. For example, on an online dating forum, or a ride-sharing service, a verifier might request evidence of a year of TE on several social media platforms in order to make their trust decision about the subject e.g. to support a proposal to meet in person (high risk). The same user wishing to make a small donation

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

or investment in a crowd-sourcing campaign online might request TE only from a single platform for a few months (lower risk). It is a matter for the subject to decide whether to disclose the requested TE (i.e. pass to the verified the relevant decryption keys) and indeed the act of declining to do so creates in itself a signal for the verifier to make their human trust decision.

This 'challenge' protocol for TE driven by the verifier is a design decision explicitly made to build in flexibility for a wide gamut of possible socio-technical interactions that may be mediated via TAPESTRY.

3.1 Privacy Attributes

In order to provide privacy to TAPESTRY users, TE is derived through a one-way hashing function that creates a compact, privacy-preserving gist of the semantic content of an activity (for example the text or image posted). TAPESTRY utilises a deep neural network (DNN) to perform this distillation in order to prevent content from being recovered from TE, yet enabling two pieces of TE to be compared to quantify the similarity of the content that generated it. The details of this process are described further within Sec. 4. Thus a subject may share evidence of an activity, such as a social media post, without providing the content of that post to the verifier. Furthermore, TE is stored within the platform in an encrypted form using a secret key held by the subject. A different key is used for each TE generated from each type of activity (Facebook photo post, Twitter text post) and is changed periodically. When a subject agrees to disclose TE to a verifier, they do so by sharing the relevant keys. Key generation and sharing, as well as the broader encryption scheme within TAPESTRY, is discussed in Sec. 5.

Since the volume of TE (e. g. spanning months or years of DP) requested of a subject is typically large, TAPESTRY creates a visual gist ('visualisation') of the TE in order to make it comprehensible to the verifier. The design of the visualisation is discussed in Sec. 6. The core information communicated via the visualisation is the coherence of the user's digital history, derived from the timestamps and similarities of the TE shared by the subject. The verifier is able to make a human decision as to whether the user is trustworthy, by having sight of this visualisation, in combination with other external factors such as social norms prevailing in their use context. At no point is an automated decision offered to the verifier as to the trustworthiness of the user, nor is the subject's decision to share TE made automatically. Rather, TAPESTRY acts as a privacy preserving conduit for the request and supply of TE.

3.2 De-centralised Trust Model

TAPESTRY is designed around a decentralised trust model, without reliance upon third-parties to vouch for the integrity and provenance of TE. This trust model is facilitated via a proof-of-work (PoW) Blockchain.

Recent changes in legislation (e.g. the European GDPR [20]) mean that users have the right to request that their personal data be deleted from any systems controlled by third-parties. In the case of TAPESTRY this meant that raw data from user activities could not be stored on-chain, as this could not be later removed. Storing only the encrypted vectors from the machine learning models, would also not comply due to the nature of the computations involved, which create an alternative digital representation of the user's behavior, thus personally identifiable information. Storing the personally identifiable information off-chain, within one or more independent data lakes (DLs), provides a method of data capture that facilitates recovering data for verification purposes, and can be deleted at the user's request.

TAPESTRY therefore uses a hybrid system of on- and off-chain storage for TE; see Fig. 1. A cloud service (of which many independently operated services are assumed to exist) maintains the DL into which the subject commits encrypted TE alongside plaintext metadata that identifies the user uniquely, along with the timestamp and type of activity. A SHA-256 hash of the encrypted TE is stored within a PoW blockchain, keyed by a hash of the metadata (computed also via SHA-256) which serves as a unique identifier to the TE record in the DL. In practice, a block committed to the PoW chain contains many such pairs. Fig. 2 summarises the interactions between the subject, the DL and the Blockchain during collection of TE. Note that the keys used to encrypt TE are different from the standard cryptographic public/private key pair used within a PoW system (here, Ethereum) to commit blocks to the Blockchain. The public key (or 'wallet address') serves also as a unique ID of the user on the sytem (c.f. pk in Sec. 5.2).

On-chain hashing enables the verifier to check the provenance of TE, prior to decrypting and visualizing that evidence for human judgement. The hash of the encrypted evidence received is compared to that stored immutably within the Blockchain. This guards against an attack via fabrication of TE by the DL provider. The PoW Blockchain is implemented via Ethereum, and a smart contract to fetch (i. e. search and retrieve) and to commit (append) data to the Blockchain is provided. Failure to verify the provenance of the data, or to decrypt the data into a parseable form (e. g. due to an invalid secret key supplied by the subject) results in an immediate rejection of the TE and strongly implies an untrustworthy interaction. Fig. 3 summarises the interactions between the verifier, subject, DL (one pictured) and the public Blockchain during sharing and verification of TE.

Given the complex security model employed to ensure secure data transmission and storage, it is plausible that user error could cause the loss of the keys, rendering their TAPESTRY data inaccessible. Users could wish to share their keys with a trusted third party key store, which could provide their keys when required or even act as a facilitator during the verification process. This optional step is analogous to sharing the private keys of cryptocurrency wallets with a centralised brokering service.

4 Extracting Trust Evidence

We now describe the process through which users' activities are distilled from explicit content and transformed into trust evidence (TE). Our approach is based on the hypothesis that people have consistent (or slowly evolving) behavior and personal interests over longitudinal time periods [58], [59]. Deviations from this normative behavior pattern indicate either a non-natural (fake) account e. g. as a vehicle for spam or online scam, or a legitimate account that has been

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

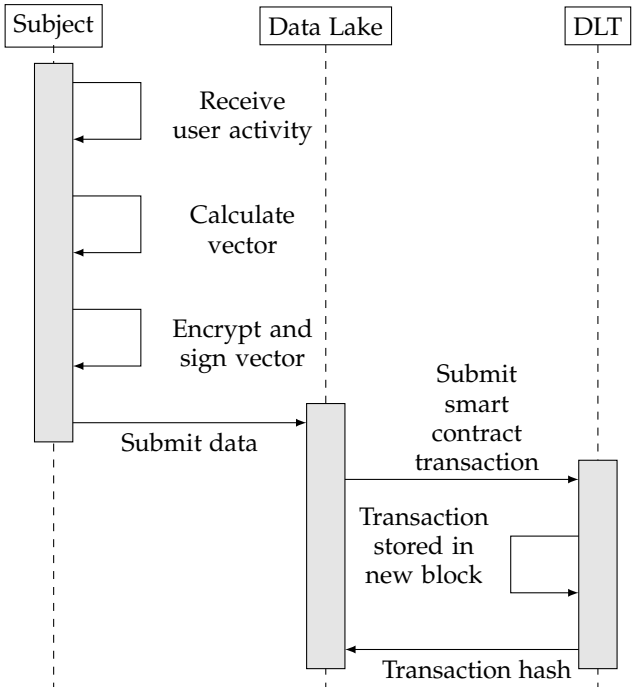

Fig. 2. Sequence diagram of the collection of trust evidence (TE) from a subject. TE is generated locally, in the form of a vector distilled from raw content via a deep neural network (DNN). The vector is encrypted, then sent to the data lake (DL) which stores the encrypted TE off-chain and records a hash of it within a new block in the PoW Blockchain (DLT).

hijacked for similar purpose resulting in abnormal characteristics in the timeline.

A central assumption of TAPESTRY is that longitudinal normative behaviour can be used as trust evidence to prove the provenance of an online identity. The use of a distributed ledger (blockchain) enables us to prove that such evidence is created contemporaneously, and cannot be 'back filled' to create an artificial history of interactions for a user ('subject'). Thus if a user behaves in a consistent manner for a considerable length of time, this is a strong signal of provenance and may influence a user (in the 'verifier' role) to trust the subject. Of course fake accounts may be created to exhibit consistent behaviour; the assumption in TAPESTRY is that creation of such behaviour over long time periods would be prohibitively expensive, and that automating such a process would leave a tell-tale behavioural signature of its own.

We tackle the problem of detecting deviation from behavioral norms as an outlier/anomaly detection task. Our goal is therefore to reduce the content of a post to a compact, real-valued vector (the TE) such that similar semantic content maps to similar TE. In this work, we present our deep neural network (DNN) based method to detect coherent/incoherent activities on the Twitter social media platform via analysis of text within a subject's posts. We make the assumption that the identity has not be compromised at the time it enrolls to TAPESTRY and begins contributing trust evidence to protect their identity. Through a combination of semantic embedding and temporal modelling, we map activity content to TE and leverage the temporal coherence of TE to help prove the provenance of a digital identity.

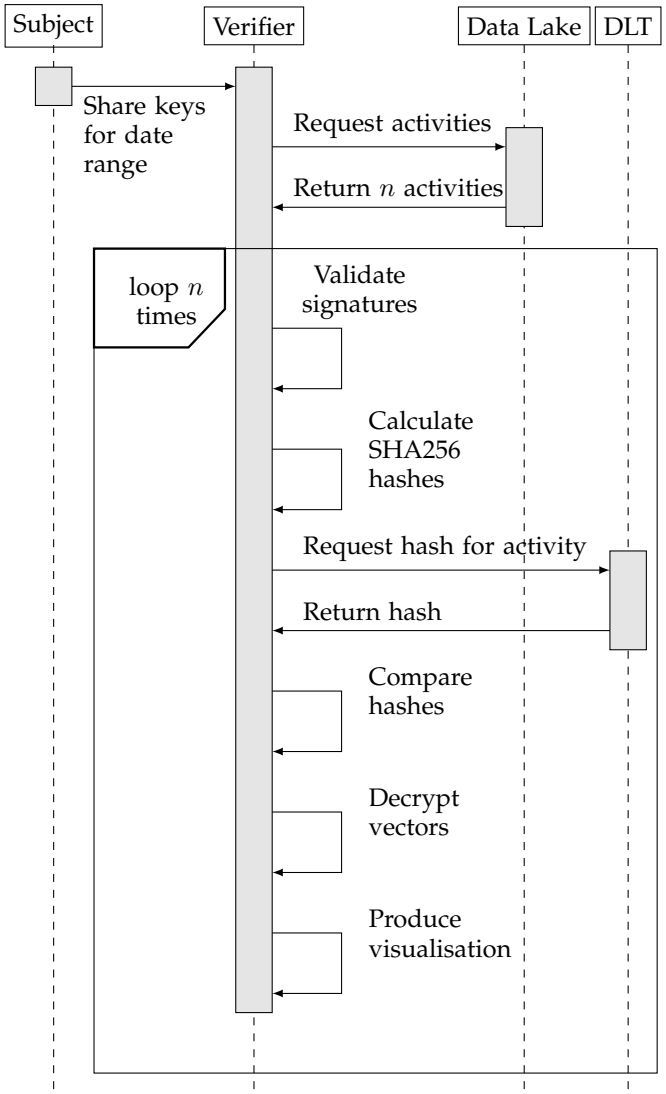

Fig. 3. Sequence diagram of the process by which a verifier determines the trustworthiness of a subject. The subject shares relevant secret keys with the verifier, enabling encrypted TE vectors to be retrieved from the data lake, verified for provenance against the Blockchain (DLT), and decrypted. The vectors are converted into a visualisation to aid the verifier's trust decision.

4.1 Data pre-processing

User-generated content (UGC) on social media is often a mixture of texts, special characters, hashtags, emojis and links. This kind of raw data is not directly suitable for machine learning methods. In natural language processing, a pre-processing step is necessary to clean the data to normalize it for the learning process. In this paper, we consider only meaningful texts and focus on topic analysis. We first remove special characters and retain only texts in a tweet. We then apply successive operations, including tokenization, stop words removal and stemming (e.g. converting words such as 'running' to 'run'), and lemmatisation (e.g. converting variants such as 'better' to simple canonical words such as 'good').

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

4.2 Topic Word Modeling

'Word embedding' is a distributed representation of words, incorporating semantic information [46] that is learned from a large corpus of text (all tweets in the collected data set in our case). 'Topical modeling' [6] extracts a distribution of words as topics, and a distribution of topics as documents. We implemented topic word embeddings, as proposed in [43], to capture contextual information in the given document. A topic word embedding is considered as a word-topic pair $< w_i, t_i >$. We considered all the tweets from one user as a document. The learned feature can enhance discrimination between words in different contexts and styles. A tweet embedding is the average of all topic word embeddings derived from the words in the tweet.

4.3 Temporal Coherence via Long-Short Term Memory

The application of Deep Learning [25] is proving highly effective in making sense of signals in computer vision [38], natural language [22] and robotics [47]. We apply a DNN to learn features for each user in a temporal window, denoted as user embedding. The user embedding is regarded as a temporal pattern of tweets in a fixed time window.

Thus user posting behaviour over a time window is characterised by the distribution of such data points in the embedding, derived from the content of posts. This distribution may be contributed to by a single user, or (if a jointly managed account) by several users; there is no difference to the algorithm which considers only the resultant distribution. When posts over a particularly timeframe deviate from this distribution significantly, then the behaviour is considered anomalous.

The Long-Short Term Memory ('LSTM') model [29] is a recurrent neural network used to model and predict time-series data. We built a sequence model to capture the coherence activities using an LSTM model and trained to extract the user's behavior norm based on their 'daily story', e.g. as played out on social media or through other online activity. We implemented a bi-directional LSTM to model the temporal coherence on a daily and weekly basis across the captured Twitter data (temporal segment). We adopted a two-layer bidirectional LSTM, followed by two fully connected layers. The input of LSTM is the topic word embedding and the output is a daily or weekly tweets embedding.

4.4 Triplet network for TE Embedding

In order to compare TE from activities over time, it is necessary to learn a metric embedding in which norms may be computed to quantify deviations in the topic word embedding over time.

Triplet DNNs have been used more broadly to learn such embeddings for information retrieval e. g. for visual search [26], [53] and we similarly perform supervised learning of the TE embedding using a triplet network strategy [56]. The objective of this network structure is to map TE within the topic word embedding to a metric embedding in which similar TE samples are pushed together and dissimilar samples pushed away from each other in the learned space. Here, similar samples are the temporal segments from one individual, and dissimilar samples are the ones from different individuals. The method proved efficient in identifying different individuals based on their temporal features, as learned by the prior LSTM step. We tested the method to detect compromised moments of an account, by randomly selecting a time step on one user's time-line feed. We then replaced the Tweets after the time point by the tweets from another user in order to simulate anomalous accounts, in order to provide negative exemplars for training.

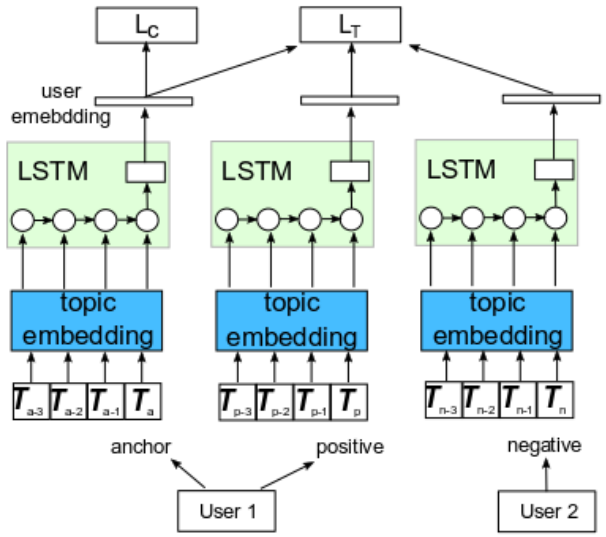

Fig. 4. Triplet LSTM architecture of our DNN projects users' activities to a high dimensional space where the generated content from the same user are close to each other, which is used as TE. Textual content is first transformed to a semantic embedding using a topic model and word embedding. Temporal embedding is learned via LSTM reflecting social media behavior over time. Initially the LSTM is trained as a classifier to discriminate between users within a training corpus (under cross entropy loss L_C, yielding a 'user' embedding) The embedding is then fine-tuned to yield the final TE embedding space via triplet training using sequences of real twitter posts (positives), and simulated fake sequences created by other users), under a triplet loss L_T.

Fig. 4 illustrates our network architecture and the end-to-end pipeline of our machine learning algorithm for TE extraction. Given a set of n users $\mathcal{U} = \{u_1, u_2, \cdots, u_n\}$, each user has a sequence of k tweets $u_i = \{T_1, T_2, \cdots, T_k\}$. The tweets are first pre-processed and projected from variable length text strings to the topic embedding space. The LSTM then learns a temporal feature of the tweets for each temporal window, denoted as user embedding, $ue_i = \{twe_1, twe_2, \cdots, twe_{k-w+1}\}$, where w is the window size. The learning strategy is a classification followed by a triplet network fine-tuning. We use a single cross-entropy loss $\mathcal{L}_C$ for pre-training and a combined classification and triplet loss $\mathcal{L}_T$ based on L^2 norm for fine-tuning [56]. In our experiments we use $n = 8000$ and $k = 800$ to train the TE embedding (c.f. Sec. 7.1).

5 Privacy and Information Security

In this section we elaborate on the data encryption and key sharing features of the TAPESTRY platform, the security properties they require and the security guarantees they provide for the service. We follow the formalism of timeline activity proofs (TAP) [17], where we have the user's trust evidence as an activity.

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

5.1 Entities and their roles

To enable formal security analysis of TAPESTRY, we model the functionalities and security assumptions of its entities, i.e. subjects, verifiers, and public ledger (data lake with PoW blockchain).

Subjects: We model subjects (i. e. users) on whom data is collected, using secret-public key pairs $(\mathsf{sk}, \mathsf{pk})$, with sk used generically to contain all secret information required by the user, i.e. signing keys and seed for derivation of the encryption key. The public key pk is used as the public identifier for the user, referred to in Sec. 3 as the user's unique ID. This key is analogous to a wallet address (e.g. in Ethereum) and we allow the user to create multiple identities, and thus, hold multiple key pairs. An assumption of our system is that a physical identity cannot transfer ownership or operation of such key pairs, thus grounding the key pair as a basis for identity in TAPESTRY.

Public ledger with external database: We consider an ideal version the public ledger that captures the functionalities of the PoW blockchain from Sec. 3, while the data lake is modelled as an external database. The public ledger is assumed trusted and cannot be corrupted; assumption that is standard for ideal formalization of the ledger. In real-world deployment, this assumption can be realized by the distributed and public nature of the PoW blockchains. The external database is also trusted not to remove/add entries, assumption easily enforced by the subject maintaining a Merkle tree root over all his encrypted TE and plaintext metadata.

Verifiers and Policies: Verifiers establish *policies* - statements over different types of activities/TE in specific intervals, that the subjects must satisfy. In TAPESTRY, all policies are subjected to a human-based decision via the visualization in Sec. 6. We model this aspect abstractly by having the policy return a Boolean value (i.e. true or false) to capture the verifier's decision. While the verifier has access to the TE or decryption keys for this TE, and is trusted to not disclose them; the verification process can only be initiated by the subject via the smart-contract in PoW blockchain, that includes an interactive proof of ownership.

5.2 Form of Trust Evidence

A subject's TE maintains a strict format: the subject's public identity pk (i. e. unique identified), the time it has been registered $\mathtt{time}$, the type of evidence $\mathtt{type}$ with any optional descriptors $[\mathtt{tags}]$, the machine learning encoding (i. e. real-valued vector) of the evidence data $\mathtt{data}$, and a digital signature σ to authenticate that it was submitted by the subject pk. When the subject submits this trust evidence, the data component is encrypted $\mathtt{cdata}$, therefore:

$$\mathrm{TE} = \langle \mathsf{pk}, \mathtt{time}, \mathtt{type}, \mathtt{cdata}, [\mathtt{tags}], \sigma \rangle .$$

Building Blocks: Our construction relies on *pseudo-random functions* (PRF) [23] and *digital signature* (DS) [15] that are *existentially unforgeable under chosen message attacks* (EUF-CMA) [24]. Additionally, we consider a *symmetric encryption scheme* (SE) with two security requirements: *indistinguishability under chosen plaintext attacks* (IND-CPA) and *wrong key detection* (WKD) [8].

Key Management: The trust evidence data is encrypted with a symmetric encryption scheme, where the encryption/decryption keys play an important part of the policy verification. Our solution is to derive unique encryption keys for finite time periods and for each kind of activity; in practice this could be as granular as a key per piece of TE. We realize this by assigning a random PRF seed s to each user, when they join our system. For TE, the encryption key ek is build as:

$$ek = \mathrm{PRF}(s, \mathrm{PRF}(s, \mathsf{pk}, \mathtt{time}), \mathtt{type}).$$

For a greater degree of granularity, we may consider counting the same type of trust evidence received at the same time duration: $\mathrm{PRF}(s, ek, \mathtt{count})$. Furthermore, this allows for a granular disclosure of encryption keys only for the trust evidence required by verifiers, without compromising the security of the other trust evidence.

5.3 Security Properties

There are two security properties that our system satisfies: *data confidentiality* that ensures the privacy of trust evidence data after the subject has submitted it to the external database associated with the ledger, and *authentication policy compliance* where verifiers are only convinced by subjects who actually satisfy verification policies. Formal definitions are available in [17].

Data Confidentiality: Intuitively, this property ensures that no information is revealed concerning the trust evidence data that the subject is submitting, just by analyzing entries in the ledger. This property is modeled by using a probabilistic polynomial-time (PPT) adversary that is required to distinguish between two private activity encodings by seeing an entry in the database that corresponds to one of them. The entries differ only on the data component, while the public key, the time, type and tags are the same for both entries. Following the formalism of TAP, we use cryptographic primitives that satisfy the security requirements of TAP: IND-CPA for the symmetric encryption scheme, and pseudo-randomness for PRF.

Authenticated Policy Compliance: This property ensures that a malicious user cannot impersonate an honest subject, or fake the existence of trust evidence in the database associated to the ledger. Therefore they cannot convince an honest verifier that they are authorized and satisfy their policy. We model this using a PPT adversary that can submit entries to the external database of a public ledger, and is considered successful if they can convince an honest verifier to accept the evidence when one of two conditions is satisfied: either the adversary impersonated an honest subject, or they provided a successful proof for a policy that they do not satisfy. Following the formalism of TAP, we use cryptographic primitives that satisfy the security requirements of TAP: WKD for the symmetric encryption scheme, EUF-CMA for the digital signature, and pseudo-randomness for PRF.

5.4 Implementation details

Our cryptographic primitives are instantiated using the implementation from the python library *pynacl*. Our PRF is instantiated with the BLAKE2b [52]. In [3], it has been shown that BLAKE2b satisfies the pseudo-randomness property

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

Fig. 5. Initial designs for the TAPESTRY visualisation prototyped with the focus group. The two designs progressed for evolution and implementation in the service, based on user feedback, were the 'slash' and the 'pie' (shown second and fourth from left on the top row). The purpose of the visualisation is to communicate the completeness of TE records over the shared time period, and the coherence of activities generating that TE i. e. to flag anomalous behaviour. In many of the initial designs, these properties were reflected by spatial coverage and use of colour respectively.

required by PRFs. Our DS uses the Ed25519 [5] implementation from [52] to instantiate the digital signature. Ed25519 offers existential unforgeability under chosen message attacks. Our SE is instantiated with the Salsa20 and Poly1305 MAC [31]. In [51] it has been shown that this construction satisfies IND-CPA. Moreover, the exact construction uses the technique from [8], and therefore also satisfies WKD.

6 Making Trust Evidence Comprehensible

TAPESTRY aims to visually communicate a gist of the completeness and coherence of a subject's TE over time, so that a verifier can make an informed choice as to whether to trust that subject. We rejected the security related motifs (e.g. ticks, padlocks) that are often used in online systems to signal the efficacy of a particular security function or domain of use; e.g. proportionate red-amber-green traffic light scales as used for food packaging to indicate nutritional content [14]; bronze, silver, gold hues often incorporated into badge, certificate or star rating symbolism. Such tropes convey trustworthiness as quantifiable and unequivocal (see [49]). All visualisations are persuasive to an extent [45]; this has serious design implications as TAPESTRY does not (visually) verify an online actor's trustworthiness but aims to support individuals in making their own judgments about in whom and what to trust. Our intention is not to make a trust decision, and then communicate that decision in a vague way. Rather, our intent (via the visualization) is to gist digital activity patterns over longitudinal time periods into a visual snapshot that will alter appearance as activity patterns are altered. The visualization acts as data to support rather than make human trust decisions, as so must avoid presenting a visual metaphor that implies a decision like a dashboard or traffic light warning system.

6.1 Prototyping of Visualisations

User focus groups and lab-based workshops with user experience (UX) designers informed early concept designs for the visualisation. Our design inspirations were broad, from Knightmare, a 1990s British TV quest gameshow for children that manifested the health status of the characters using pixelated computer graphics, to more conventional information and data visualisation practices. From this we then produced 12 initial designs we called 'snowflake'; 'slash'; 'radiance'; 'pie'; 'T-bar' (referring to TAPESTRY); 'tiles'; 'pixel face'; 'Picasso'; 'T'; 'shield'; 'eye' and 'pixel head' (see Fig. 5). We then rejected designs that could not depict sufficient granularity of either the completeness or coherence of TE over time. We also rejected designs that did not readily scale down (i.e. for viewing on a small screen) e. g. 'pixel head' inspired by Knightmare, the most anthropomorphic of the designs. We also rejected 'eye' as evocative of a surveillance system. We explored use of colour and tone, both to enable additional granularity of visual representation of the shared TE and with regard to colour's culturally situated function that could invite potentially unintended meanings for some users. Additionally, in terms of interpreting completeness of TE – within the research team it became apparent that the computer scientists associated lighter tone with more TE while designers interpreted white areas of a design as an absence of TE within a given time period. With these constraints in mind, we selected to use the idioms of 'slash' and 'pie' as the preliminary visualisations for further development.

6.2 Evolved Visualisation Designs

The final designs for the 'pie' and 'slash' TE visualisations are shown in Fig. 6. The choice of two designs for the visualisation reflect two use cases for deployment of TAPESTRY, evaluated in Sec. 7.

Pie is based on a simple dial that lends itself to representing temporality and accumulation of DP over days, weeks, months etc, across concentric circles, as though accumulated TE is moving towards the core of the pie. This design is used for the interpersonal trust case in which users are required to make trust decisions on an *a priori* unknown online business or service. We report on the efficacy of the visualisation in this context within the video games crowdfunding experiment of Sec. 7.2.

Slash meanwhile was intended to communicate introspective trust, where users can check TE derived from their own DP (i. e. act as both subject and verifier) to determine whether their online accounts have been hacked (Sec. 7.1.). This required a design that could visually detail sudden dissonance within an otherwise relatively uniform pattern of DP as generated over time. A visual design analogy would be a ladder in hosiery, or a dropped stitch in knitwear; these draw the eye, despite their small scale, to solicit a feeling of unease in the user to invite further investigation.

Both these visualisations necessitate some initial explanation and guidance [45], though their intended meaning will require learning only once [12]. This is addressed via a tutorial during the initial user sign-up to the TAPESTRY service.

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

Fig. 6. Developed visualisations deployed in TAPESTRY. Left: Pie communicates interpersonal trust e. g. to help users determine whether to trust an unknown business or service online. Concentric rings of the pie correspond to different granularities of time, and shading is used to communicate coherence and volume of activity within the time period corresponding to each segment. Right: Slash communicates introspective trust e. g. to help users determine if their social media account has been compromised. Each slash corresponds to a period of time, with backslashes indicating anomalous (outlier) TE during that period. Shading is used to communicate volume of TE during that period, as with 'pie'. Six instances of the visualisations are displayed ranging from complete and coherent, to sparse TE over the time period requested by the verifier.

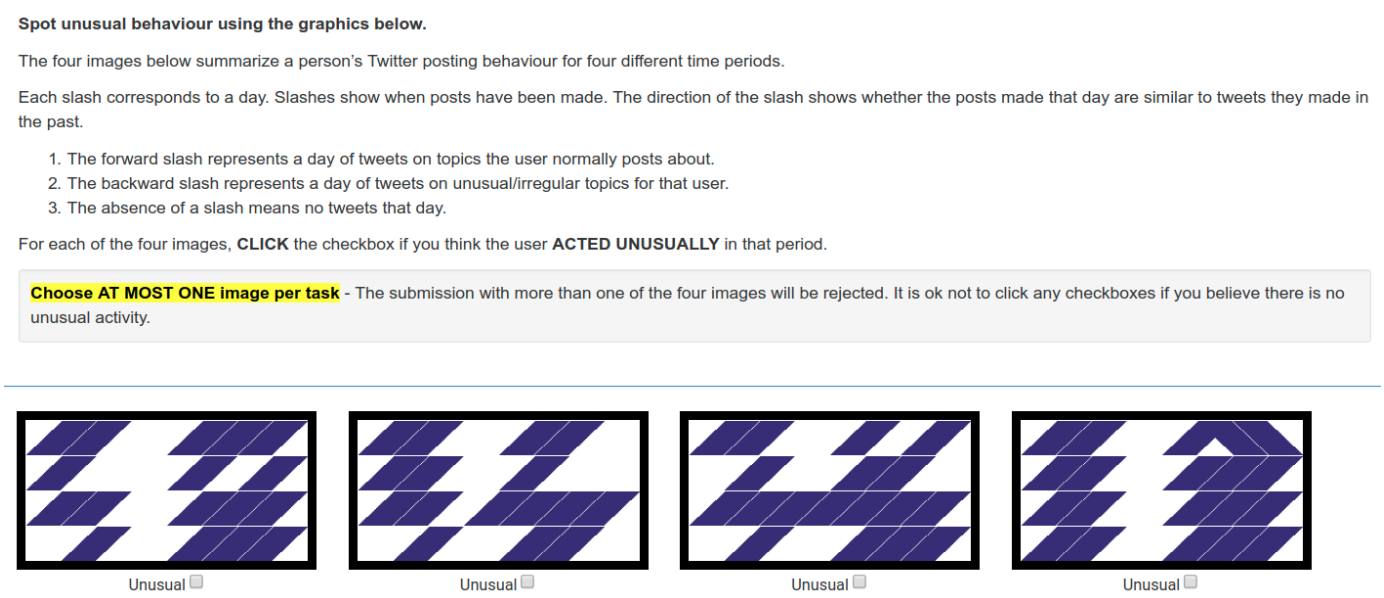

Fig. 7. Mechanical Turk experiment evaluating the efficacy of the 'slash' visualisation for anomaly detection (introspective trust). MTurk workers were presented with a series of four visualisations, each summarizing four weeks of trust evidence. Based on the series, they were asked which (if any) of the visualisations appeared anomalous relative to the others, and thus implied that unusual activity was present in the social media feed.

7 Experimental Evaluation

We evaluate the TAPESTRY service in two contexts. First, introspective trust, which we evaluate in the context of social media account hijacking; a user ('verifier') must determine whether their social account has been compromised due to anomalous posting behaviour. Second, interpersonal trust in which a user ('verifier') must determine the trustworthiness of another ('subject') online. We evaluate this in the domain of rewards-based crowdfunding, where parties are typically unknown to one another initially, and a trust judgement is fundamental to deciding whether to invest. In both experiments we have focused upon the Twitter platform, deriving trust evidence from text in user posts.

7.1 Introspective Trust: Anomaly Detection

We first evaluate the efficacy of our proposed DNN approach (Sec. 4) for extracting trust evidence from social activities (text-based Twitter posts). We evaluate the ability of the approach to discriminate between the behaviour of different users, and its ability to detect anomalies within the social media feed of individual users. The experiments were conducted using a public dataset of Twitter posts ('tweets') gathered by Li et al. [41] initially comprising 50 million tweets for 140,000 users. In our experiments, we study social media footprints over longitudinal time periods and clean the data by removing users with fewer than 800 tweets in their timeline feeds. The remaining 8000 users form the basis for our experiments.

Evaluating Trust Evidence (TE) Embedding

We justify our choice of a LSTM to learn a temporal model for TE, via comparative evaluation against two state-of-the-art DNN architectures; RNN and GRU. We evaluate all three architectures as a user classification problem: the networks are trained using 80% of the tweets of $n = [500, 8000]$

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

	Precision			Recall			F1-score		
f_h	0.1	0.2	0.3	0.1	0.2	0.3	0.1	0.2	0.3
OCSVM	0.43	0.46	0.48	0.44	0.46	0.48	0.43	0.46	0.47
IF	0.71	0.62	0.57	0.75	0.64	0.59	0.73	0.63	0.58
LOF	0.12	0.24	0.34	0.13	0.25	0.37	0.12	0.24	0.35
Ours	**0.93**	**0.94**	**0.95**	**0.98**	**0.91**	**0.92**	**0.95**	**0.91**	**0.94**

TABLE 1
Evaluating the ability of our embedding to perform anomaly detection. Results compare our proposed approach to anomaly detection within the learned embedding to two common baselines.

users and tested on the remainder. Accuracy is measured as the number of times the system correctly identifies the user among the n possibilities. Fig. 8 shows the result of classification accuracy on all the three models as n increase. The LSTM model outperforms the other two models in most cases.

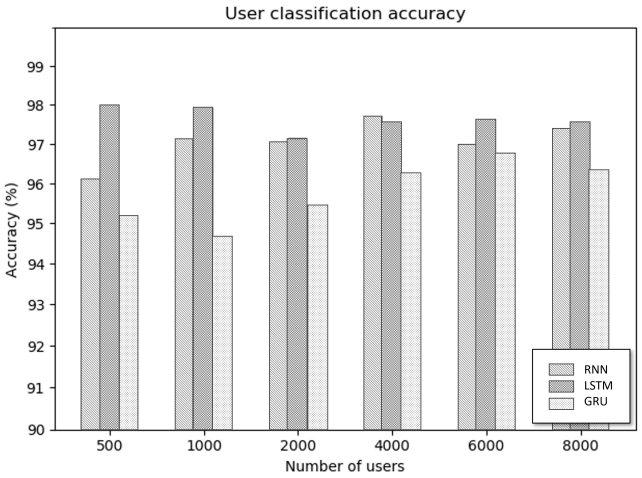

Fig. 8. Evaluating user discrimination without our DNN learned embedding for trust evidence. Our temporal modelling approach (based on LSTM) is compared to RNN and GRU sequence models. The experiment is run for $n = [500, 8000]$ corpus of users and accuracy measured as the $\%$ of users correctly identified based on 20% of their TE.

Evaluating Anomaly Detection

We compare the efficacy of our learned embedding at detecting anomalies within a single user's history of TE. For this experiment we train the model on 80% of users, and test on the remaining 20%. We compare several approaches to detecting anomalies within the test partition:

1) **One-class SVM (OCSVM)** [55] computes a non-linear boundary in a higher dimension space using kernel method for data projection. . This method allows for only positive data as 'one class'.
2) **Isolation Forest (IF)** [42] evaluates the isolation degree of each data point using a random forest. This algorithm focuses on separating the outliers from the data points.
3) **Local outliers factor (LOF)** [7] is a distance-based method using Euclidean distance considering the density of neighborhood information.
4) **Proposed method.** We applied a distance based method of outlier detection within the proposed user embedding. We compute the distance between each data point and the distribution of data within the embedding corresponding from the first fraction of user's activity, which we consider representative of the users' behavioural norm. We consider distances larger than a selected threshold as outliers (see subsec 7.1 for detail).

We simulate hijacking actions on users' Twitter stream to evaluate anomaly detection. Given a fraction (f_h) hijacking length, we randomly replace a fraction of tweets in each user's feed by another user's content. These replacements create unusual behaviors that deviate from user's activity norm. Three baselines were selected to be representative a spectrum of statistical approaches commonly used in anomaly detection (SVM classification, random forests, density estimation). Table 1 quantifies the performance of our proposed approach (in terms of both precision and recall) against these baselines. Our proposed method – LSTM – outperforms all the other methods regardless of the duration of the period the account is hijacked for.

Synthesizing the Visualization (Slash)

The slash visualization comprises four rows of glyphs; each glyph is a slash or a backslash representing coherent of incoherent behaviour versus the user's historic norm. Each row comprises 10 glyphs, collectively representing a week of Twitter activity and thus the visualization comprising four rows (Fig. 7) comprises approximately a month of social media activity. The user's historic norm is modelled as a Gaussian mixture model (GMM) fitted to the distribution of historic tweets, each of which maps to a real-valued vector within the TE embedding (Sec. 4). In our user study the distribution was built using the previous month of activity data from Twitter. In order to determine if a glyph (representing approximately 10% of a week's activity) should indicate coherence or not, we average the Mahalanobis distances of all tweets in that period to GMM. If there is no activity during the time period then no glyph is produced (a white 'gap' is left in the visualization).

Comprehensibility of the visualisation

In order to evaluate if our proposed visualisation is understandable and helpful for non-expert users to detect potential anomalies, we crowd-source evaluation on the Amazon Mechanical Turk (MTurk) platform. MTurk was used to recruit 92 non-technical participants with day to day experience using social media sites. Fig. 7 depicts a representative questionnaire. We provide a series of four visualisations of a user's activities, sampled at regular intervals across their TE history. Each visualisation contains four weeks' of activities. Normal behaviors are represented by forward slashes, while backward slashes represent unusual behaviors detected by our proposed method in Section 4. A blank position means no activities in that time-stamp. Based on the series presented to them, the MTurk workers are asked which (if any) of the visualisations appear anomalous relative to the others, and thus imply that unusual activity is present in the social media feed. The evaluation draws on the online activity of 1000 users in our dataset. We excluded data from users who had less than 4 weeks' of activities in their timeline, leaving us with 537 users whose activities are represented by visualisations, and used in the evaluation as MTurk tasks. Each task is assigned to 5 different workers. A task is correctly detected if most (i.e. 3 out of 5) of the

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

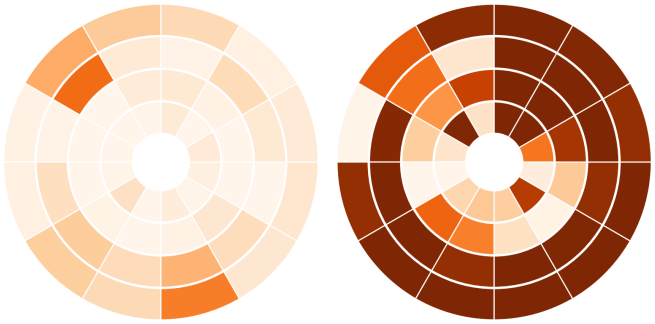

Fig. 9. Visualisation of two crowdfunders' digital footprints: strong provenance (left) and limited provenance (right). The TE requested by the 'verifier' (participant) was fixed to a month with concentric circles represent daily tweet activity (inner) to four weeks of activity (outer).

MTurk workers make the correct decision. On average, each worker spent around 1 minute per task, including reading the instructions and submitting their decisions.

7.2 Interpersonal Trust: Crowdfunding

We evaluate the performance of TAPESTRY in terms of benefit to the human decision-making process when establishing the trustworthiness of an individual online. An initial account of this experiment was given in the IEEE Vizsec workshop as a short paper [62].

Mock Crowdfunding Campaign

We designed a user study that involved making a decision on whether to invest in a (mock) crowdfunding campaign for a video game start-up - a sector heavily reliant upon such funding. In this context, a user (the 'verifier') is invited to make an investment by another user (the 'subject') without prior knowledge of that subject. Our video game was offered for investment by eight crowdfunders on a mock platform: these crowdfunders were a mix of four real games industry professionals and four fake identities that had been created two months prior to the study. We commissioned an experienced video game narrative writer who mocked-up a pitch for a new video game and an associated crowdfunding campaign. We had gained the consent of the games developers to use their real profiles in the campaign. Meanwhile the writer produced four fake profiles based on their knowledge of the gaming industry. We created fake Twitter accounts for the fake profiles and continually tweeted relevant game and entrepreneur-related comments for two months prior to the study to provide historic context. All eight crowdfunders had one campaign web page hosted on a password protected micro-site, which included a description of the game (constant across all candidates), a short biography for each profile and a link to their Twitter account. The creation date of the real and fake Twitter accounts was obfuscated.

Synthesizing the Visualization (Pie)

The Pie visualization comprises four concentric rings, each segmented into twelve segments that are shaded to indicate coherence of social media activity. Each ring represents activity within a different time period, from inner ring to outer: one day, one week, one fortnight (two weeks), four weeks (approximatley one month) – all from the current date. The segments represent one twelfth of that time period and are shaded light to dark to indicate the degree of coherence (lighter) or incoherence (darker, or absent) of social posting behaviour versus the user's historic norm. The user's historic norm is modelled as a Gaussian mixture model (GMM) fitted to the distribution of the previous month of tweets mapped to a real-valued vector within the TE embedding (Sec. 4). When determining the darkness of a segment, we shade inversely proportional to the average of the Mahalanobis distances of all tweets in that period to GMM. If there is no activity during the time period then the segment is absent (a white 'gap' is left in the visualization).

Experimental setting

The study was run in a closed workshop with 10 participants recruited from the University campus population. The 10 participants were aged 25-40 and gender balanced. The recruited participants were non-technical, and had no prior experience of crowd-funding nor specialist knowledge of the games industry. After a briefing on the TAPESTRY service, participants were invited to read the crowdfunding campaigns, browse the background description and biographies and invest a hypothetical $1000 'TAPESTRY currency' between the eight campaigns. We randomly split the participant group into two groups of five; only one group was provided with TAPESTRY visualisations on the mock crowdfunding site (Fig. 9). Participants could use the mock site or wider resources on the Internet to help them to make decisions to allocate the money. The study lasted 35 minutes; participants were asked to make one decision every 5 minutes using the knowledge they gleaned from their full use of Internet resources.

Experiment results

We evaluated the participants' performance based on their investment results, comparing the amounts invested in real and fake profiles for both groups. We consider investment in a fake profile (i. e. scammer) a bad investment. Fig. 10 shows that the accuracy of the investment results correlates to the time taken in background research; the more participants gathered information from their searches on the Internet, the more accurately they made their investment. Given the time limit, the TAPESTRY group used the visualisation tool to quickly understand the games developers' Twitter identity, speeding up their search to establish legitimacy. We can conclude that although participants reached similar, correct decisions (in terms of discriminating their investment between genuine and fake developers) the time-to-task was considerable shorter (approx. by half) for TAPESTRY users.

8 Conclusion

We presented TAPESTRY, a novel decentralised service that enables users to determine the provenance of online identity from their digital personhood (DP) in order to make better decisions on who to trust online. We applied the TAPESTRY service to two tasks; determining the trustworthiness of another unknown individual (interpersonal trust), and determining the integrity of one's own social media feed (introspective trust). We used machine learning techniques to extract trust evidence from social media activities in a privacy preserving manner, and a proof-of-work Blockchain to store hashes of that evidence in order to underwrite its provenance. Our service enabled users to then selectively

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

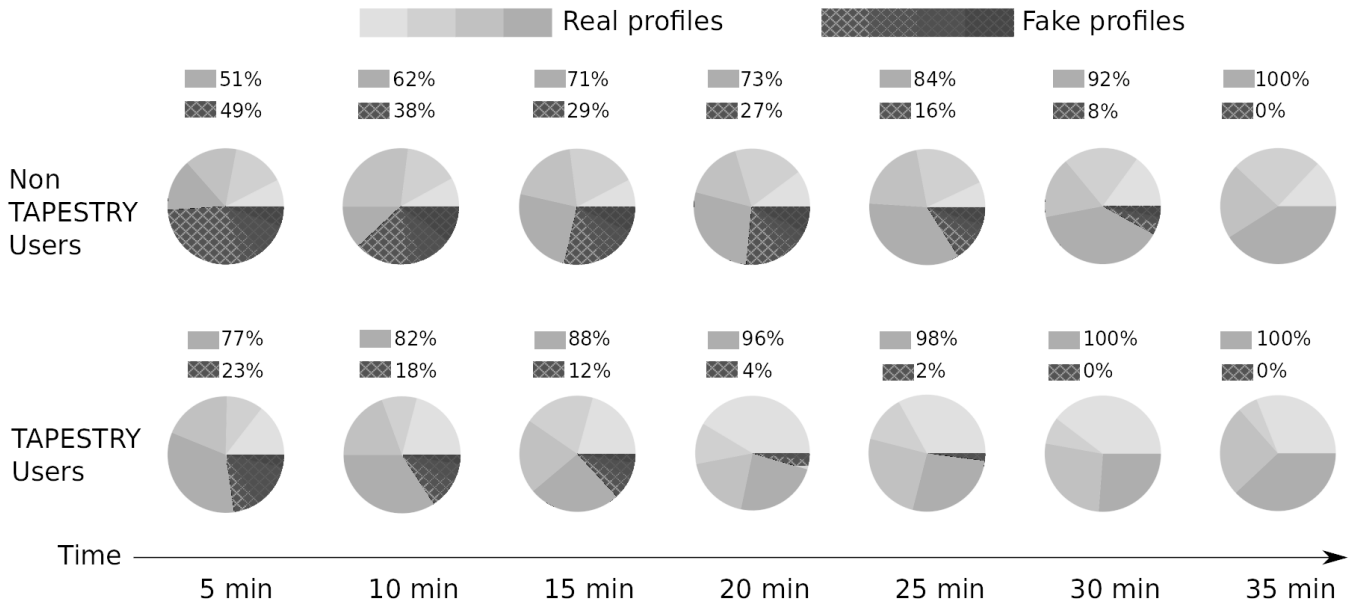

Fig. 10. Participants' investment performance in the mock crowdfunding experiment: top row is the control group without TAPESTRY visualisation; bottom row with TAPESTRY. The users with access to TAPESTRY achieved similar or better accuracy in detecting fake profiles, but did so at least twice as quickly.

disclose trust evidence to one another in order to prove the provenance of their identity. To improve comprehension of the high volumes of evidence shared between users, we designed visualization techniques to summarise that evidence. We evaluated the end-to-end system using a mocked up crowdfunding exercise run in a user workshop, and showed that TAPESTRY enables people to make accurate trust decision faster than the control group who lacked access to the service. We evaluated the end-to-end system for anomaly detection and showed that TAPESTRY enabled users to detect anomalies in a social media feed with accuracy of $\sim 94\%$.

Currently TAPESTRY is a prototype and future work will explore at-scale deployment beyond workshop settings. At scale, it will become necessary to run multiple data lake services, with users distributed across different lakes. This will add value to the PoW Blockchain which will then be maintained across multiple lakes. At this stage, further characterization of the performance of the TAPESTRY Blockchain should be undertaken. In addition, the studies in this paper focus on just a single social media modality (text, from Twitter posts) and a scaled-up system would explore multiple platforms and modalities. Nevertheless, we do not believe at-scale deployment of TAPESTRY is necessary to demonstrate the value in our hybrid on- and off-chain architecture for identity provenance, and the novel machine learning and visualization techniques developed for the service.

TAPESTRY is underpinned by the requirement that users who wish to interact online maintain an active digital personhood (DP) from which trust evidence may be derived. Whilst we do require that users' maintain a DP, we place no requirement on the regularity with which users maintain it (e.g. post). Afterall, if a user irregularly posts to social media, and that is their behavioural norm, then this is the pattern of behaviour that other users (verifiers) would learn to expect from that user – or if monitoring one's own social media feed, the pattern one expects to see in the visualization for that feed. Nevertheless the project raises broader societal questions around digital inclusion should be considered where individual may (e.g. due to age, or personal choice) opt not to maintain a DP; one might leverage IoT or other digital interactions in place of social media (as proposed here) in such circumstances.

ACKNOWLEDGMENTS

TAPESTRY is funded by EPSRC Grant Ref: EP/N02799X/1 under the UKRI Digital Economy Programme.

REFERENCES

[1] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews by network effects. In *Seventh international AAAI conference on weblogs and social media*, 2013.
[2] M. Andreesen. Why bitcoin matters. *New York Times*, 2014.
[3] J. Aumasson, S. Neves, Z. Wilcox-O'Hearn, and C. Winnerlein. BLAKE2: simpler, smaller, fast as MD5. In M. J. J. Jr., M. E. Locasto, P. Mohassel, and R. Safavi-Naini, editors, *Applied Cryptography and Network Security - 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings*, pages 119–135, 2013. LNCS 7954.
[4] A. Barth, D. Boneh, and B. Waters. Privacy in Encrypted Content Distribution Using Private Broadcast Encryption. In *Financial Cryptography and Data Security'06*, pages 52–64, 2006. LNCS 4107.
[5] D. J. Bernstein, N. Duif, T. Lange, P. Schwabe, and B. Yang. High-speed high-security signatures. *J. Cryptographic Engineering*, 2(2):77–89, 2012.
[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.
[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.
[8] R. Canetti, Y. T. Kalai, M. Varia, and D. Wichs. On Symmetric Encryption and Point Obfuscation. In *TCC'10*, pages 52–71, 2010. LNCS 5978.
[9] R. Chalapathy, E. Toth, and S. Chawla. Group anomaly detection using deep generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 173–189. Springer, 2018.

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

[10] C. Cheshire. Online trust, trustworthiness, or assurance? *Daedalus*, 140(4):49–58, 2011.
[11] R. Claudini. The science of persuasion. *Scientific American*, 284(2):76–81, 2001.
[12] A. Cooper, R. Reimann, and D. Cronin. *About Face 3: The Essentials of Interaction Design 3rd ed.* Wiley, Indianapolis IN., 2017.
[13] L. A. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *IEEE Communications Magazine*, 47(12):94–101, 2009.
[14] P. Desai, M. E. Levine, D. J. Albers, and L. Mamykina. Pictures worth a thousand words: Reflections on visualizing personal blood glucose forecasts for individuals with type 2 diabetes. In *Proc. CHI Conf. on Human Factors in Computing Systems*, New York, 2018. ACM.
[15] W. Diffie and M. E. Hellman. New directions in cryptography. *IEEE Trans. Information Theory*, 22(6):644–654, 1976.
[16] R. B. D.M. Rousseau, S.B. Sitkin and C. Camerer. Not so different after all a cross-discipline view of trust. *Academy of Management Review*, 23:393–404, 1998.
[17] C. C. Dragan and M. Manulis. Bootstrapping online trust: Timeline activity proofs. In *DPM & CBT- ESORICS 2018*, pages 242–259, 2018. LNCS 11025.
[18] P. Dunphy and F. A. P. Petitcolas. A first look at identity management schemes on the blockchain. *IEEE Security Privacy*, 16(4):20–29, July 2018.
[19] C. Elsden, A. Manohar, J. Briggs, M. Harding, C. Speed, and J. Vines. Making sense of blockchain applications: A typology for hci. In *Proc. CHI Conf. on Human Factors in Computing Systems*, pages 458:1–458:14, New York, 2018. ACM.
[20] EU. General data protection regulation 679, 2016.
[21] G. Farnadi, J. Tang, M. De Cock, and M.-F. Moens. User profiling through deep multimodal fusion. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 2018.
[22] Y. Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.
[23] O. Goldreich, S. Goldwasser, and S. Micali. How to construct random functions. *J. ACM*, 33(4):792–807, 1986.
[24] S. Goldwasser, S. Micali, and A. C. Yao. Strong signature schemes. In D. S. Johnson, R. Fagin, M. L. Fredman, D. Harel, R. M. Karp, N. A. Lynch, C. H. Papadimitriou, R. L. Rivest, W. L. Ruzzo, and J. I. Seiferas, editors, *Proceedings of the 15th Annual ACM Symposium on Theory of Computing, 25-27 April, 1983, Boston, Massachusetts, USA*, pages 431–439, 1983.
[25] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.
[26] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *Proc. ECCV*, pages 241–257, 2016.
[27] F. Günther, M. Manulis, and T. Strufe. Cryptographic Treatment of Private User Profiles. In *Financial Cryptography Workshops'11*, pages 40–54, 2011. LNCS 7126.
[28] J. Hancock and J. Guillory. Deception with technology. In *The handbook of the psychology of communication technology*, page 270–289. Wiley-Blackwell, 2015.
[29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
[30] L. Hornuf and A. Schwienbacher. Portal design and funding dynamics. *SSRN Electronic Journal*, 2015.
[31] R. Housley. Using ChaCha20-Poly1305 Authenticated Encryption in the Cryptographic Message Syntax (CMS). *RFC 8103*, 2017.
[32] H. Jones and W. Moncur. The role of psychology in understanding online trust. In *Psychological and Behavioral Examinations in Cyber Security*, page 109–132. IGI Global, 2018.
[33] H. S. Jones and W. Moncur. A mixed-methods approach to understanding funder trust and due diligence processes in online crowdfunding investment. *Trans. Soc. Comput.*, 3(1), Feb. 2020.
[34] U. Kang, L. Akoglu, and D. H. Chau. Big graph mining for the web and social media: algorithms, anomaly detection, and applications. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 677–678, 2014.
[35] K. Kim and S. Viswanathan. The experts in the crowd: The role of reputable investors in a crowdfunding market, 2014.
[36] A. D. K.Orzech, W. Moncur and D. Trujillo-Pisanty. Opportunities and challenges of the digital lifespan: views of service providers and citizens in the uk. *Information, Communication Society*, 21(1):14–29, 2018.
[37] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, pages 5802–5805, 2013.
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
[39] V. Kuppuswamy and B. Bayus. Crowdfunding creative ideas: The dynamics of project backers in kickstarter. *SSRN Electronic Journal*, 2013.
[40] R. Lambiotte and M. Kosinski. Tracking the digital footprints of personality. *Proceedings of the IEEE*, 102(12):1934–1939, 2014.
[41] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *KDD*, pages 1023–1031, 2012.
[42] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.
[43] Y. Liu, Z. Liu, T.-S. Chua, and M. Sun. Topical word embeddings. In *AAAI*, pages 2418–2424, 2015.
[44] B. W. M. Tzanetakis, G. Kamphausen and R. von Laufenberg. The transparency paradox. building trust, resolving disputes and optimising logistics on conventional and online drugs markets. *International Journal of Drug Policy*, 35:58–68, 2016.
[45] M.Correll. Ethical dimensions of visualization research. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2019.
[46] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.
[47] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
[48] S. Nevin, R. Gleasure, P. O'Reilly, J. Feller, S. Li, and J. Cristoforo. Social identity and social media activities in equity crowdfunding. In *Proceedings of the 13th International Symposium on Open Collaboration*. ACM, 2017.
[49] J. Nurse, I. Agrafiotis, M. Goldsmith, S. Creese, and K. Lamberts. Two sides of the coin: measuring and communicating the trustworthiness of online information. *J. of Trust Management*, 1(5), 2014.
[50] C. Phua, V. Lee, K. Smith, and R. Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.
[51] G. Procter. A Security Analysis of the Composition of ChaCha20 and Poly1305. *IACR Cryptology ePrint Archive*, 2014:613, 2014.
[52] PyNaCL. https://pypi.org/project/PyNaCl/.
[53] F. Radenović, G. Tolias, and O. Chum. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In *Proc. ECCV*, pages 3–20, 2016.
[54] E. Rocco. Trust breaks down in electronic contexts but can be repaired by some initial face-toface contact. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, volume 23, page 393–404. ACM, 1998.
[55] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.
[56] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
[57] S. Vismara. Equity retention and social network theory in equity crowdfunding. *Small Business Economics*, 46(4):579–590, 2016.
[58] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *USENIX Security Symposium*, pages 223–238, 2014.
[59] C. Wang and B. Yang. Composite behavioral modeling for identity theft detection in online social networks. *arXiv preprint arXiv:1801.06825*, 2018.
[60] C. C. Whitehill. The financial cost of fraud 2018. Technical report, University of Portsmouth, 2018.
[61] R. V. Yampolskiy and V. Govindaraju. Behavioural biometrics: a survey and classification. *International Journal of Biometrics*, 1(1):81–113, 2008.
[62] Y. Yang, J. Collomosse, A. Manohar, J. Briggs, and J. Steane. Tapestry: Visualizing interwoven identities for trust provenance. In *VizSec 2018 -15th IEEE Symposium on Visualization for Cyber Security, 22nd October 2018, Berlin,Germany*, 2018.
[63] R. Yu, X. He, and Y. Liu. Glad: group anomaly detection in social media analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(2):18, 2015.
[64] B. Y. Zhao, L. Huang, J. Stribling, S. Rhea, A. Joseph, and J. Kubiatowicz. Tapestry: a resilient global-scale overlay for service deployment. *IEEE J. on Selected Areas in Communications*, 22(1):41–52, January 2004.

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TSC.2020.2993081, IEEE Transactions on Services Computing

Yifan Yang is a Research Fellow at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. His research interests include but not limited to machine learning, image understanding and knowledge representation and reasoning.

Daniel Cooper is a Research Software Developer at the Centre for Vision Speech and Signal Processing (CVSSP), University of Surrey. He manages the University of Surrey DLT testbed infrastructure, as well as the design and development of software, integrating AI technologies with DLT.

John Collomosse is a Professor at the Centre for Vision Speech and Signal Processing, University of Surrey. His interests are in the fusion of Artificial Intelligence and Distributed Ledger Technology (DLT), and he directs the Surrey Blockchain testbed comprising several UKRI funded projects in this area.

Constantin Cătălin Drăgan is a Research Fellow at the Surrey Centre for Cyber Security (SCCS) and Department of Computer Science, University of Surrey, UK. His research interests are in areas of applied cryptography, security analysis, formal methods, and privacy. Previously, he worked as Research Fellow for CNRS & INRIA, France.

Mark Manulis is Deputy Director of Surrey Centre for Cyber Security (SCCS) and Senior Lecturer in the Department of Computer Science, University of Surrey, UK. His research interests are in applied cryptography, network security and privacy.

Jamie Steane is Associate Professor At Northumbria School of Design, Newcastle upon Tyne, UK. His design research lies at the intersection of design, business and digital technology and has involved the commercial development of early interactive products and services for the creative industries, financial and educational sectors.

Arthi Manohar is a researcher and lecturer in the Department of Design, Brunel University London. Her research interests include participatory design, co-design, user-centered design and Human Computer Interaction.

Jo Briggs is Associate Professor at Northumbria School of Design, Newcastle upon Tyne, UK. Her research concerns designing tools for safer and enjoyable online interaction and investigations into and through the creative collaborative economy. She leads on design for a number of interdisciplinary projects and publishes on design and socio-technical subjects.

Helen Jones is a lecturer and researcher in Experimental Social Psychology at the University of Central Lancashire, UK. Her primary research interest lies in understanding how people behave online, in particular focusing on the decision-making processes that can leave them vulnerable to cyber security threats.

Wendy Moncur is Interdisciplinary Professor of Digital Living at the University of Dundee. She leads the Living Digital group, which focuses on human factors in cybersecurity, online identity, personal data and delivering personal agency to users.

1939-1374 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Northumbria University Library. Downloaded on May 11,2020 at 10:31:42 UTC from IEEE Xplore. Restrictions apply.