The Irrefutable History of You: Distributed Ledgers and Semantics for Ubiquitous Personal Ratings

Allan Third and John Domingue

Knowledge Media Institute, Open University, Milton Keynes, MK7 6AA, UK
{allan.third,john.domingue}@open.ac.uk

Abstract. A recurring theme in the science-fiction series *Black Mirror* is the consequence for society of an over-focus on social networking. The episode *Nosedive* imagines a future in which every public interaction a person has is rated by the other parties, and every aspect of ones life depends on the overall *rating* computed from these. In this paper, we show how such a scenario is already technically possible using existing technologies such as *distributed ledgers*, and discuss means by which the negative possibilities may be ameliorated using semantic approaches.

1 Introduction

The television drama *Black Mirror* [4] focuses, in each standalone episode, on the potential personal and social consequences of the use of technology, basing its plots on forms of technology which can be imagined as at least partially plausible extensions of what is available or in use today. The societies depicted are usually in some sense dystopian, although not always (*San Junipero* [5] being a notable counterexample). A recurring theme across a number of episodes is the idea of an over-focus on social networking.

The episode *Nosedive* [8] depicts a world in which every public interaction with another person can be rated, between zero and five stars, by means of a smartphone – after any interaction, pointing the phone's camera at the person one wishes to rate brings up a picture of that person and an interface to rate them. Ratings can also be given for social media posts. Ratings are aggregated for each person, so that it is also possible for anyone to view someone's aggregate rating, again, by holding up their phone or looking at an online profile. The ubiquity of ratings has developed huge social importance, with employment, personal and social consequences – a person might qualify, or not, to rent a home, or get a job, or enter a social venue based on their rating. The society shown is one of constant effort to maintain or improve ones rating, and the plot of the episode follows how one woman's attempts to significantly improve her own rating go wrong, and through a series of unfortunate events, backfire, leading to her rating nosediving from just over four to nearly zero over a short period of time.

We first show that the scenario of *Nosedive* can already be implemented using technologies available now, before discussing possible technical approaches which could minimise the negative social consequences of a hypothetical ubiquitous adoption of such a rating system.

2 Distributed Ledgers

An emerging area of research in the Web concerns the use of *distributed ledgers*, based on the *blockchain* data structure, to serve as immutable trustworthy record stores in environments where trust is lacking. Originally developed to underpin the Bitcoin cryptocurrency [12], blockchains are being experimented with for a number of use cases, including verifiable educational certification [9, 14, 17], data integrity and access [1, 18] and Internet-of-Things applications [6, 19]. The design of a blockchain is such that the integrity of its contents are guaranteed by a large network of financially and organisationally independent nodes, and the cost of editing a record is prohibitively high, meaning that once entered onto a blockchain, a record is effectively immutable and irrefutable.

A distributed ledger is based on a blockchain: a timestamped sequential series of records shared *in toto* across a network of nodes. There is no central or authoritative node, and every node has the potential to add new records to the ledger. New records are added by consensus: anyone who wishes to add data to a node submits a cryptographically-signed request to do so. Each node competes for the right to add ("mine") a new block of transactions to the blockchain, with that right awarded by consensus among all the nodes. The precise method of consensus varies between different types of blockchain. The important feature of any consensus mechanism is that there should be a cost involved in attempting to mine blocks, and a reward for the successful node, to encourage good behaviour among nodes. The winning node selects a set of pending record transactions and groups them into a *block*, which is added to the chain. Requiring a consensus between a large network of (financially, and organisationally) unrelated nodes in order to add any data to the chain guarantees that spurious data cannot be inserted by a malicious agent, and, due to the fact that modifying an older record requires the entire chain following the modified record be rewritten as if from scratch, the records entered in a blockchain are effectively immutable.

Distributed ledger platforms such as Ethereum [20] add the ability to have executable code embedded in a blockchain, in so-called "smart contracts". Because the code of a smart contract is stored on the blockchain like any other record, it is immutable and traceable to the account of the person who deployed it to the platform. This means that smart contracts can be relied on as not having been tampered with. If the source code of a contract is also made available, it is possible to verify that the deployed version is genuinely created from the given source. An open source smart contract, then, can be trusted as a genuine implementation of what its source code describes, which allows everyone who interacts with the contract to verify that it behaves in the way that is expected

of it. Smart contracts are therefore usable to mediate transactions between users on a distributed ledger platform in a trustworthy way.

Records on a distributed ledger are public, and, as stated earlier, all nodes on the network have a complete copy of the entire chain. This does not mean, however, that everything must be readable by everyone; records can contain encrypted data, and, with a suitable cryptographic key management infrastructure, the data owner can implement a "selective visibility" system, only permitting chosen others to view the full or partial contents of a record.

3 Implementing the *Nosedive* rating system

We see two main technical requirements to implement the *Nosedive* rating system. To simplify the discussion, let us say that person A wishes to give person B a rating of n stars, where n is a natural number between 0 and 5 in relation to a face-to-face interaction I between A and B. Person C, at a later date, wants to view this rating's value.

1. the ability for A's mobile device to recognise B's proximity and provide A with the ability to select B's name and apply the rating.
2. Ratings must be unforgeable and incapable of being tampered with, and always attributable to both A and B in their respective roles in the rating transaction.

In the episode, A simply needs to hold her device up facing B rather than making a selection on the screen to achieve 1; potentially we can imagine advanced facial recognition in addition to the other techniques described here enabling this. And 2 is never explicitly stated; however, we must assume that it applies in order for the system to carry the weight in people's lives that it is shown to do.

To implement 1, it is possible to take advantage of the widespread availability of accurate GPS devices in modern smartphones. If each phone registered its current location in a public database, it would be possible for A's phone to use its own location to query for nearby phones, and then, from the returned list of devices, query each in turn, via some presumably standardised interface, for a user profile, containing, e.g., B's name, picture, cryptographic public key and "rating account" details. To ensure as best as possible that ratings of B were tied to B's actual behaviour (and not, for example, the behaviour of someone who has stolen B's phone), it may be necessary to require some biometric identification of B. For non-proximate ratings, profiles can be shared via social media.

2 can be developed using a distributed ledger with smart contracts. If A and B both have accounts on a distributed ledger, publicised via their user profiles, then it would easily be possible for A to submit a numerical rating to B's "rating" smart contract via a cryptographically-signed request. If C reads B's profile, then he can query B's rating contract to view the rating n. A suitable public key infrastructure would support trustworthy attribution of the rating to A and prevent forgery, and the properties of the distributed ledger would ensure

that ratings could not be tampered with once made – collectively making the ratings effectively irrefutable. In fact, we are currently experimenting with the scenario of distributed-ledger-based reputation transactions ([9]) using the idea of "reputation tokens" to be used in education as a means of recognising and accrediting soft skills that traditionally are not adequately covered by traditional assessment and accreditation.

Signed and immutably timestamped records on the ledger would provide, therefore, an effectively irrefutable record of all rated interactions in a person's history with the system. While it would be possible, and perhaps simpler in some ways, to implement a system using a centralised database without a ledger, one could not then be certain that ratings could not be tampered with by those with access to the database.

The technologies required are already developed and in use, and it would not be technically difficult to build the *Nosedive* system using them.

4 Potential for abuse

It should be noted at the outset that there are already currently-obtaining situations in which people voluntarily take part in systems in which they are rated by other users, either explicitly or implicitly. Every user of the auction site eBay [7], for example, has a rating determined by other users reflecting their status as a "good citizen" of the site. Forum/discussion sites which implement a "Like" button on contributions can show the total number of "likes" a user has received on the site. (The tabletop games community site Boardgamegeek [3], for example, does this with "thumbs", its "like" system.) In a somewhat more sophisticated way, online dating sites provide rankings of users, although these rankings are context-sensitive and depend on the user viewing them, in terms of how well users match the viewers. The algorithms computing the rankings are generally proprietary; it seems reasonable to assume that they are in some way learned from similar users.

The bulk of existing research into online rating systems has focused on online sales systems or discussion forums such as these. Desiderata for an online reputation system are suggested in [21], and include mechanisms to prevent or discourage users from changing their identities to "discard" a bad rating, and to limit the effect of "memory" in ratings - to prevent historical low ratings from dragging an average down even when a person's current ratings are consistently high. One of the desiderata is also that ratings from high-ranking users be weighted more than those of low-ranking ones – an idea which we discuss below in section 5.1.

It is clear, however, that the disastrous outcome for the protagonist in *Nosedive* is not far-fetched; a system for ubiquitous ratings of people, for any reason, and with profound social consequences, is ripe for abuse. We envisage this abuse in the form of malicious, trivial or thoughtless low ratings directly, as shown in the episode, as well as in over-interpretation of low ratings in inappropriate contexts. In existing ranking systems, [16] showed that negative rankings have

a disproportionate effect on reputation compared to positive rankings. While it may well be appropriate to refuse to buy from an eBay seller with a low rating, it is clearly not appropriate to refuse to hire someone on the basis of, for example, their social life. Even supposing that one accepted the concept of rating each public interaction according to perceived merit, there is no *a priori* reason to believe that "fair" ratings are achievable. People have been shown to assess in others in a discriminatory manner based on existing prejudices in otherwise identical circumstances (see, e.g., [13]). We may therefore reasonably expect that groups disadvantaged on the basis of gender, ethnic origin, sexuality, and so on, would become disadvantaged in terms of ratings. Indeed, in *Nosedive*, those who are shown to be experience the most significant consequences as a result of low ratings are members of existing disadvantaged groups, by gender or race.. The protagonist's brother, however (white, male and hinted to be heterosexual) is shown to be less invested in the need for a high rating, and it seems plausible that ones investment in ratings in general would vary in inverse proportion to societal privilege.

5 Minimising the effects of a *Nosedive* rating system

Let us suppose that some social or economic pressure leads to the widespread adoption of such a technology. In itself this appears to be unlikely, for reasons including the scenario which is the central plot of *Nosedive*, but for the sake of argument, let us assume that it takes place. And let us assume that such a system could be enforced, perhaps by economic pressures, to be close to universal, so that everyone had access to the relevant hardware and software to take part, and the options for "opting out" were limited to non-existent. (In practice, we assume that none of these are achievable, particularly the latter two, and that such a system has so many manifest and serious ethical and social problems with it that widespread adoption would be resisted very very strongly, sufficiently to undermine its implementation at all.) But if it did, what measures could be put in place to protect users from malicious or punitive ratings?

5.1 Automatic or manual moderation

The most obvious, and on the surface, likely to be effective, approach would be if B had the power to accept or reject ratings at will, but of course, to do so would undermine the concept of a rating and would lead to the vast majority of, if not all, people having only the maximum rating. Even a slight modification, in which ratings are accepted or rejected, but the recipient does not know the value of the rating at the time of acceptance, would still tend to lead to overall high average ratings, as people would generally choose to accept ratings for interactions which they felt had gone well.

Take into account the rating of the rater Low-rated people's ratings do not count as much as those of high-rated people, i.e., A's ratings are weighted by her own rating.

This approach might seem good on paper, but is likely to be harmful in practice. Given the hypothesis that the distribution of high and low ratings would come to reflect existing power structures in society, there is the potential for this approach simply to end up entrenching disadvantage. An episode of the sitcom *Community* [2] with a similar theme to *Nosedive* envisaged this precise scenario, with the outcome being a very highly stratified society with little social movement.

Take into account the rating history of raters. The ratings of someone who frequently rates people low are weighted less than the ratings of others.

This is also subject to being gamed. All a malicious person needs to do is to give high ratings to interactions which are *not* important to them in order to maintain rating influence on those who do. It also would only affect persistently low-rating people, and would have no effect on the situation where B receives many low ratings from a large number of otherwise typically-rating people. (This may well have been the case at a number of points in the episode, such as when the protagonist receives a lot of low ratings because she's standing on the highway in traffic, or when she takes over the microphone at the wedding.)

This would also be relevant to other forms of gaming of the rating system. For example, family and friend groups might cooperate to increase their members' ratings, while antagonistic social groups could do the opposite. Analysis of rating history could identify patterns such as this, and allow for normalisation of ratings to reflect them.

Require mutual ratings If only A rates B in relation to I, nothing happens to B's rating until and unless B also rates A in relation to I.

This would provide B with some control and the ability to avoid ratings which are predicted to be bad. However, bearing in mind we are discussing a hypothetical situation where society as a whole has decided to adopt ubiquitous rating, it seems less than likely that this level of control would be accepted.

Mutual agreement would be one of the mechanisms by which erroneous ratings could be corrected (for example, in the case of misidentification of a person).

Meta-rating Nearby people are randomly selected to "rate a rating", with low-ranked ratings being withdrawn or reversed.

This approach has the potential to make a difference, and has been implemented with positive effects in discussion systems – a notable example being the Slashdot technology news site [15], where every contribution in a discussion can be rated by other users, and whose "metamoderation" system has been shown to succeed in maintaining civility in discussions [10]. This method could also be used to address erroneous ratings.

Potential difficulties with this method are the overhead of "meta-rating" – would people be willing to expend the extra effort to do it? – and the observation that meta-ratings are just as likely to be affected by conscious and unconscious bias as the ratings themselves.

Apply a cost to giving ratings If there were a cost to rating someone, or a limit on the number of ratings which could be given (within a particular time period, for example), then it may serve to deter frivolous ratings. If the cost were financial, of course, this would again serve to entrench existing status. If simply a limit in the number of ratings, there is a chance that malicious ratings would be less likely.

5.2 Semantic approaches

Contextual ratings Rather than a semantics-free number, ratings are instead applied to a particular category of interaction or behaviour, e.g., "this rating is for a customer in a retail purchasing interaction". [11] argues that ratings should always be interpreted relative to their original context.

This allows fine-grained information to be conveyed. Specifically, it allows for more subtle *interpretation* of ratings. So in hiring someone, an employer will not (indeed, should not) be concerned with what that person is like as a wedding guest, and can therefore exclude all ratings related to irrelevant transactions. More specifically, an aggregate rating could be computed not across all interactions, but across all interactions *of specified types*, such as all financial interactions. With more sophisticated interfaces and possibly smarter systems, the context could be extended to support the notion of *evidence* for a rating to be recorded alongside the rating itself.

For contextual ratings to function, there would need to be an ontology of human interaction describing the categorical relationships between interaction types (such as that a retail transaction is a financial interaction, for example). Manual modelling, however, is only likely to get so far, given the range and complexity of human interactions which take place. Fine-grained categorisations could be crowdsourced, with, perhaps, some central editorial approval or moderation process to avoid discriminatory or offensive categories.

Reasoning could be carried out on semantic representations to derive new categories or specialised ratings – this has potential for positive or negative consequences. For example, it might then be possible for malicious users to simulate discriminatory categories indirectly by deriving them. Positive possibilities include, among others, the ability to identify patterns in ratings received, which might help to limit or avoid negative effects. (The same approach would apply if it were also possible to find out the ratings that someone had *given* too, which would improve transparency in the process regardless.)

Purely semantic ratings The idea of contextual ratings could be taken one step further, and the idea of numerical ratings could be dropped entirely. So instead of "B is rated 4", or "B is rated 4 in relation to I, which is a retail interaction", it would be "B is rated "helpful" in relation to I, which is a retail interaction". Models would need to be developed for relevant attributes, such as "helpful" which could be associated with particular interactions, or interaction types, and with sufficient ontological semantics, reasoning could be carried out as with numbers. Such contextual semantic ratings would be more fine-grained still, and

offer more flexibility, while perhaps helping people to avoid the habit of over-interpreting numbers. Aggregation too need not be purely numerical; aggregate categories using vague qualifiers, such as "mostly harmless", could be computed.

Selective visibility of ratings It could even be possible, using smart contracts, to enforce restrictions on ratings that are taken into account; if C has to specify up-front the reason for reading B's rating, with B's acceptance of the reason, there could be hardcoded, crowdsourced, or legally mandated sets of interaction types which may be considered in computing the rating. Potentially the user could have the ability to override the request and add further limits onto the considered interaction types. However, this latter ability seems likely to be subject to the same social pressures as requiring approval or mutual rating before a new rating has effect.

Two-way personalisation Combining semantic representation of contextual rating data with selective visibility gives the potential for all parties, A, B and C, to provide semantic descriptions of preferences and inputs which can be used for fine-grained negotiation of requests and permissions when it comes to rating, being rated and viewing other ratings, from which one would expect good practices to emerge from privacy-conscious users which could be adopted by those less skilled when it comes to technology and privacy protections.

6 Conclusions

We have described how a ubiquitous personal rating system, as shown in the *Black Mirror* episode *Nosedive*, could be implemented using technologies available today, backed up by smart-contract-enabled distributed ledger platforms which ensure the availability of effectively immutable histories of effectively irrefutable records of ratings. We have speculatively discussed potential moderation techniques and Semantic Web technologies which might aim to limit the negative consequences of such a system.

None of the speculated measures do more than mitigate the possibilities of abuse, and it is hard to see how anything other than avoiding the system at all could prevent it entirely. Currently-used personal rating systems are limited to very specific contexts, such as eBay, with no consequences beyond those contexts, but a ubiquitous system is inevitably subject to many forms of abuse.

While we believe it is highly unlikely that society as a whole would adopt a *Nosedive* system, there have of course been historical and recent instances of societies choosing options not in the best interest of the majority of its members. Expecting a technological solution to scenarios such as this is naïve; technology can be designed to support safe choices or minimise dangerous ones, but ultimately, it can rarely, if ever, enforce them. if there were a significant danger that *Nosedive*-style ratings were to become a ubiquitous phenomenon, the most effective approach would likely be to tackle the economic and social pressures leading to this danger.

References

1. Azaria, A., Ekblaw, A., Vieira, T., Lippman, A.: Medrec: Using blockchain for medical data access and permission management. In: Open and Big Data (OBD), International Conference on. pp. 25–30. IEEE (2016)
2. Blum, J., Deay, P., Saccardo, T., Kolb, C., Ridley, R., Padrick, M., Diego, D., Roller, M., Harmon, D.: App development and condiments. Community S5(E8) (2014)
3. BoardGameGeek: https://boardgamegeek.com (Jul 2017)
4. Brooker, C.: Black Mirror. Zeppotron (2011)
5. Brooker, C.: San Junipero. Black Mirror S3(E4) (2016)
6. Christidis, K., Devetsikiotis, M.: Blockchains and smart contracts for the internet of things. IEEE Access 4, 2292–2303 (2016)
7. eBay: https://ebay.com (Jul 2017)
8. Jones, R., Schur, M., Brooker, C.: Nosedive. Black Mirror S3(E1) (2016)
9. KMi: http://blockchain.open.ac.uk (Jan 2017)
10. Lampe, C., Zube, P., Lee, J., Park, C.H., Johnston, E.: Crowdsourcing civility: A natural experiment examining the effects of distributed moderation in online forums. Government Information Quarterly 31(2), 317 – 326 (2014), http://www.sciencedirect.com/science/article/pii/S0740624X14000021
11. Mui, L., Mohtashemi, M., Ang, C., Szolovits, P., Halberstadt, A.: Ratings in distributed systems: A bayesian approach. In: Proceedings of the Workshop on Information Technologies and Systems (WITS). pp. 1–7 (2001)
12. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)
13. Oreopoulos, P.: Why do skilled immigrants struggle in the labor market? a field experiment with thirteen thousand resumes. American Economic Journal. Economic Policy 3(4), 148 (2011)
14. Sharples, M., Domingue, J.: The blockchain and kudos: A distributed system for educational record, reputation and reward. In: European Conference on Technology Enhanced Learning. pp. 490–496. Springer (2016)
15. Slashdot: http://slashdot.org (Jul 2017)
16. Standifird, S.S.: Reputation and e-commerce: ebay auctions and the asymmetrical impact of positive and negative ratings. Journal of management 27(3), 279–295 (2001)
17. Third, A., Domingue, J., Bachler, M., Quick, K.: Blockchains and the Web position paper. In: W3C Workshop on Distributed Ledgers on the Web (2016)
18. Third, A., Tiddi, I., Bastianelli, E., Valentine, C., Domingue, J.: Towards the temporal streaming of graph data on distributed ledgers. In: 2nd International Workshop on Linked Data and Distributed Ledgers, Supplementary Proceedings of the 14th Extended Semantic Web Conference (forthcoming 2017)
19. Valentine, C.: GreenDATA. http://projects.kmi.open.ac.uk/greendata (2016)
20. Wood, G.: Ethereum: A secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper (2014)
21. Zacharia, G., Maes, P.: Trust management through reputation mechanisms. Applied Artificial Intelligence 14(9), 881–907 (2000)