The Universal Trust Machine: A survey on the Web3 path towards realisation.

Rohan Madhwal
5568412
TU Delft
Delft, Netherlands
R.Madhwal@student.tudelft.nl

***Abstract*—Since the dawn of human civilization, trust has been the core challenge of social organization. Trust functions to reduce the effort individuals or groups would spend in constantly monitoring the actions of others in order to verify their assertions, thus facilitating cooperation by allowing the creation of groups which function with reduced complexity. In modern societies, this trust is provided by large centralized institutions. Specifically in the case of the Internet, Big Tech companies like Facebook, Google etc control who can read, publish and interact with content. However, as recent events have shown, allowing for-profit corporations to harness so much power and act as gatekeepers to content comes with a litany of problems. While so far ecosystems of trust on the Internet could only be feasibly created by large big tech institutions, Web3 is an emerging future vision of the Internet whose proponents aim to create an ecosystem where trust is generated without centralised actors. They attempt to do so using decentralised technology and trust generated purely from mathematical primitives. This survey attempts to explore this elusive goal of Web3 to create a "Universal Trust Machine", owned by both nobody and everybody. In order to do so, we discuss the attempts at the decades-old problem of generating trust without an intermediary and explain the lessons learnt from that research for Web3 development. Further, we uncover contemporary techniques used to achieve this goal and establish Web3's progress on its path towards realisation.**

I. Introduction

Humans in a society rely on trust in every stage of their life, in every action they perform. Children trust that their parents will nurture and guide them, adults trust that their family and loved ones won't deceive them. When crossing the street on a zebra crossing, we trust that motorists will obey the traffic laws, when buying items at the market, we trust in the quality of the goods being provided to us. Regardless of whether one believes that society is a function of divine order or of a social contract, trust between its members is the very fabric of its organising foundation. [1]

More generally, consider an agent, such as a human or a robot, who is required to use limited agency to navigate and take actions in a world with limited direct information available to it at any given moment. In such a world, trust is an important social heuristic that allows the agent to make wagers on the predictive benevolence of other agents. [2]

Hardin defines trust as "encapsulated interest", since it facilitates peaceful and stable social relations that form the basis of collective behavior and productive cooperation. Hobbes argues that the natural state of humans is nasty and brutish, however, trust helps to convert that into something peaceful and efficient. In his book "A treatise of human nature", David Hume discusses the importance of trust to the functioning of a society. According to Luhmann, trust effectively reduces complexity and risks, allowing for coordination with increased performance. [3]. This is easy to understand intuitively since trusting individuals and groups reduces the effort one would spend in constantly monitoring the actions of others in order to verify their assertions. It is easy to conclude that a society without a notion of trust would find it hard to function effectively, or to exist at all. [4]

The growth of human civilization from small-scale hunter-gatherer societies to thriving economies of nation states is testament to the benefits provided by the growth of trust and cooperation inside societies. However, history reminds us that the requirement of trust for facilitating cooperation also leads to the growth of large centralized institutions since these institutions historically provided the best defense in economic transactions against the untrustworthy. [5]

While trust might be fundamental to cooperation in a society, underlying every social transaction is the desire to further one's personal gain by abusing the trust of an unsuspecting opponent and defecting against the expected trustworthy action. [1] For example, in a transaction where a merchant pre-pays a farmer for their produce at the end of the year, the farmer may be tempted to keep the payment and not provide the promised crops, or provide crops of a lower quality than was agreed upon.

According to Margart Levi, "good defenses make good neighbors". Hence, the need for such defenses in economic transactions necessitated institutional bases of reaching agreement and resolving disputes that might result from them. Institutions that were able to provide third party enforcement in a transaction were hence able to ensure personal security and the security of the transaction. Thus, they were able to encourage cooperation and grow immensely as a result of their importance in doing so. [5]

However, allowing profit driven institutions to amass so much power comes with its own set of problems. The financial crisis of 2008 which was primarily attributed to failure of trusted institutions such as banks and other financial institutions has led to a growing distrust in such institutions [6]. This was most notably witnessed by the recent growth of blockchain

technology and adoption of cryptocurrencies such as Bitcoin and Ethereum as a decentralised alternatives to large financial institutions.

Even though the Internet was built on distributed protocols, large scale cooperation was similarly consolidated around a few centralised services where social trust was created and enforced by large profit driven institutions [7]. Specifically, in two key functions of the web, web-publishing and discovery of content, technological institutions such as Google, Meta and Twitter slowly became curators and gatekeepers for the information being published on the Internet and people who were allowed to interact with it. As a result of this, the platforms accrued the power to control and own a large share of the information published and consumed on the Internet.

Recently however, abuses of information and communication technology by such institutions for surveillance, spreading of disinformation and coercion of the public have come to light. Notable examples include Google's deepening involvement with Egypt's repressive government and Twitter enabling the Chinese government to promote disinformation on the repression of Uighurs. [8]

Such propensity of Big Tech organisations to abuse their ecosystems of trust for their own profit through privacy violations and misinformation is leading to a shift in the general attitude towards large centralised information platforms. The requirement of a large centralised authority or platform owner to maintain and enforce trust in sociotechnical systems is increasingly being viewed more as a hindrance rather than a help. [8]

A growing alternative to the existing model of the platform driven Internet is the idea of Web3 which is motivated by the idea of using decentralised technologies such as blockchain. It is hard to exactly define Web3 since there is a lack of consensus even among researchers on what the idea of Web3 means. In section II we attempt to clearly define what Web3 refers to in the context of the paper. On a high level, Web3 can be thought of as an ecosystem of applications which aims to generate trust purely through decentralised technology and mathematical primitives. Thus, Web3 aims to be a "Universal Trust Machine", eliminating the need for profit driven organisations and allowing for the creation of a "commons" [9] where everybody is free to publish, read, react, and interact with content. However, as seen so far, fostering cooperation in a community with the presence of bad actors is not a trivial problem and realising the dream of Web3 is a long, arduous path.

The problem of cooperation has been studied in the field of game theory such as in the Prisoner's Dilemma and analysing studies in this field for developing systems where the best course of actions for neighbours is to cooperate for mutual good motivates how decentralised systems may be able to function effectively. We believe that lessons

Plethora of research also exists on models and mathematical primitives for generating trust in decentralised systems, most notably, reputation systems have gained prominence as a way to create safe and trustable communities in decentralised networks. [10]

This survey attempts to explore such mechanisms for generating trust in Web3. In section II we explain the motivation behind Web3 and the technologies and movement behind its origin. After this, we discuss problems one faces when designing a decentralised system which fosters long term cooperation in section III. Then, in section IV we discuss some principles in the work of Evolution of Cooperation which help motivate how long term cooperation could come about naturally. Finally, in section V we discuss some existing work in the field of Reputation Mechanisms for decentralised systems.

II. History and Background

A. Decentralisation and Decentralised Networks

Decentralisation is not a novel concept and has been prevalent in research even outside the sciences. In the social sciences, it boasts a 200 year history and has been a popular concept across multiple disciplines. Examples include concepts such as subsidiarity, democracy, liberty and equality in political science, systems theory and self determination in management and decision science, fiscal decentralisation in economics. [11]

In technology, the concepts of technological decentralisation have been evolving for over half a century [11]. A popular example of a decentralised IT movement is the open source software movement which represents a radical retake on copyright law and involves developing and sharing software in a decentralized and collaborative way, relying on peer review and community production.

The importance and the success of this movement is demonstrated by the domination of multiple areas of software by open source projects. Popular examples are the the open source Apache projects which dominates the market of server software over commercial alternatives from Microsoft, Sun etc and the Linux operating system which has seen popular use being embedded in a range of devices from mobile phones, recording devices to large scale servers in data centers. [12]

The concept of a "decentralised network" was first coined by Paul Baran, one of the inventors of packet switching. In general, networks can be classified as two components, "star" or centralized and "grid"/"mesh" or distributed. In a star/centralized network, all nodes are connected to a single node, hence, each participant needs to go through a central component to interact with each other. While in a distributed network on the other hand, there is no such central node and each node can communicate with each other without going through a centralised point. In practice, a combination of these components is used to form a network, Baran called such a mixed network "decentralised" because there was no single, central point of failure [13]. Fig. 1 demonstrates these networks visually.

In contemporary modern literature, the term decentralised network is used to refer to networks where the technology, content and infrastructure on the network is controlled by

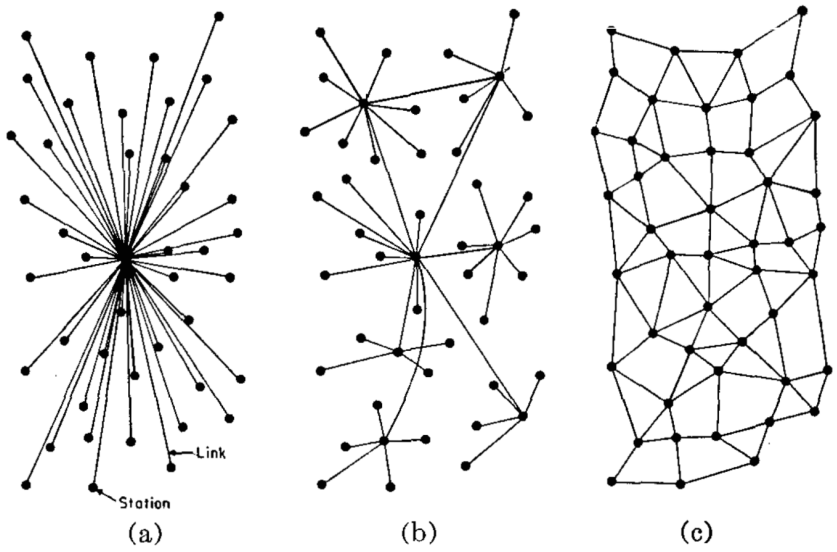

Fig. 1. a) Centralised network b) Decentralised network c) Distributed network [13]

participants and contributors rather than large central platforms. This control is manifested in various ways, such as participants controlling parts of the infrastructure like servers and routers, collaborators owning data in their own private data silos which is queried by the network during discovery, participants possessing the autonomy to decide the operational details of the network, what content needs to be publicised and what needs to be deleted etc. [7] In this context, a popular example of a centralised network would be Twitter which owns all the content that users publish on it, while an example of a decentralised network is Tribler, a peer to peer file sharing system which improves upon the BitTorrent protocol which enables users to share content with keyword search and boasts a reputation-management system to encourage collaboration. [14]

Over the past decade, decentralised networks have received a reinvigorated interest due to the emergence in popularity of cryptocurrencies such as Bitcoin and Ethereum. In his whitepaper proposing Bitcoin, Satoshi Nakamoto proposed a novel decentralised peer-to-peer network protocol which facilitates an electronic payment system [15]. The popularity of these cryptocurrencies has also resulted in explosion of blockchain and decentralised technologies and projects. Proponents and developers of these technologies wish to see a shift in the publishing and discovery of content and information over the Internet away from a few profit-driven Big Tech corporations and into the hands of the users who generate them, guaranteeing the privacy of their data and also ensuring that everyone has a fair and equal voice.

For example, “78 days”, a collaborative project between the Starling Lab and Reuters uses decentralised ledgers to preserve historical data important to humanity. The goal of the project is to curb misinformation. It achieves so by ensuring the integrity and authenticity of the information as it captured and stored using a system called Content Authenticity Initiative. It also uses a storage system built on blockchain called Filecoin that requires data providers to prove that they are holding the authentic data and not a tampered version. Most importantly it ensures that the contributors of the information have a way to maintain their creation of the content through the records stored with the data. [16]

B. Web3

The term “Web2.0” was first coined by Tim O’Reilly in 2007 to describe an Internet where platforms enabled users to publish, consume and interact with content, and with each other. [17] It was supposed to expand upon the first iteration of the Internet or “Web1.0” which largely consisted of static pages meant only to display information. So while “Web1.0” was ”the read web”, “Web2.0” aimed to be the “the read-write web” (coined by Richard McManus in 2003).

Critics of Web2.0, such as the inventor of the World Wide Web, Tim Berners-Lee feel that Web2.0 failed to achieve the vision of the Internet as a secure, decentralised exchange of public and private data, with users’ data being increasingly stored in corporate data silos. Instead, to guarantee security of their data, they want users to own their own data. [18]

The term “Web3.0” was coined by Polkadot and Ethereum co-founder Gavin Wood in 2014, he used it to describe an Internet that is decentralised, open and transparent. [19]

The current Web3 movement aims to transform the platform oriented Web2.0 Internet into a decentralised web ecosystem which: 1) avoids monopoly of content discovery and propagation by large centralised actors 2) prevents the spread of misinformation and fake news 3) provides its users the ability to create, exchange and react to information in a secure, private and free manner 4) supports immersive web development [11]

Liu et al [20] define Web3 as a movement which agnostic of any specific overarching applications or underlying infras-

tructures will usher in "an era of computing where the critical computing of applications is verifiable", that is, an application that conforms to the idea of Web3 is one where all stakeholders are able to verify the execution of the application based on predetermined terms without the presence of an intermediary.

Packy McCormick defines Web3 as "the internet owned by the builders and users, orchestrated with tokens" [21] Defining Web3 with its key property being user ownership is a common approach taken by a majority of research papers on the topic. Hence, Web3 is positioned as the "read, write, own" web. While Web2.0 was a frontend revolution that allowed users to create and interact with created content online, Web3 is instead a backend revolution which aims to change how the created content is stored. Instead of keeping data on centralised data silos, Web3 aims to provide data storage to users in a distributed manner in a way that users can own and monetise the content they created. Thus, it aims for the disintermediation of existing parties such as large big tech companies in data governance. [22]

III. Threats to long term cooperation in a decentralised network

In order to enable the dream of Web3, it is fundamental to be able to create a commons with communities of users interacting with each other through decentralised networks, free to read, publish and interact with content. However, two broad classes of threats make creating long term cooperation in decentralised networks a non-trivial task: Social and Infrastructural threats. In the following sections we briefly cover these threats and establish why they pose a problem to cooperation.

A. Social Threats

As seen by recent events, the rise of populist movements stands to be the biggest threat to the state of democracy worldwide. Many observers, especially journalists have suggested that the rise and spread of these movements has been massively aided through social media. [23] While social media can be a powerful tool for spreading information, when left unregulated, it can also lead to multiple social issues which greatly threaten long term cooperation. Some of these issues are:

1) Echo Chambers and Polarisation: "Echo Chambers" are used to describe the mechanism by which people on sociotechnical platforms are exposed to large or exclusively pro-attitudinal communication. Such grouping of like minded people on social networks ('homophily') is believed to arise from preferential connection to like minded individuals when creating/breaking bonds and also from peer influence which results in connected individuals growing more similar. [24] The presence of an Echo Chamber could support populist messages that support rejection of expertise and reasoned debate among different views and lead to the emphasis of popularity of people or ideas over substance of their views. Therefore, Echo Chambers can lead to an insulation of users from the truth and even more perniciously, to be exposed to fake news.

In their study on Echo Chambers in the context of COVID-19 discussions on Twitter, Jiang et al [25] found strong evidence of political echo chambers on the topic on both ends of the political spectrum, but particularly so in the right-winged community. They found that tweets by right leaning users were almost exclusively retweeted by users who were also right leaning. Further, from random walk simulations, it was found that information in right leaning bubbles rarely travelled out of that bubble, forming a "small, yet intense political bubble". In another study on Climate Change discussions on Twitter, Williams et al "found a high degree of polarisation in attitudes, consistent with self selection bias" [24]

Studies have suggested that echo chambers could lead to polarisation of users and thus to users retreating into like-minded networks [26], which creates segmentation in networks and thus poses a large challenge to long term cooperation.

2) Inequality and Social Divide: While the idea of a digital democracy is appealing, it is hampered by findings of socioeconomic inequality which prevent usage of the platforms by certain stratas of society. Beyond inability to access platforms, it is possible that members of society lack the skills to express their views or consume information that is being shared by other members. [27]

A lack of participation by different members of society could lead to the propagation of biased views or misinformation against the underrepresented members. Thus, it constitutes a credible threat to long term cooperation.

However, diffusion theories predict inequality at the outset of any innovation which is narrowed as time progresses and adoption rate spreads.

B. Infrastructural Threats

In section I, we motivated why trust is fundamental to achieving cooperation inside communities. Since in a Web3 application based on a decentralised network there are no third parties for enforcing trust, before using a service to cooperate with other nodes in the network, users look for 'assurance' that the other party is trustable. This is especially true for applications that depend on blockchain technology due to the immutable nature of transactions making it incredibly hard to punish bad actors. [28] Therefore, in addition to social problems, there are also several infrastructural problems stemming from the presence of bad actors who wish to abuse the trust of their neighbours for their own benefits makes the problem of achieving long term cooperation in a decentralised network a non-trivial task. A system that will be able to achieve the stated dreams of Web3 should be effectively able to tackle these problems, below is a brief description of a few of these problems:

1) Free riding: In order to encourage successful long term cooperation, it is important that enough peers are providing sufficient resources for the system to become large and truly useful. In the absence of a third party monitoring each user, it is possible that some users stop contributing and only consume resources being generated by other users. *Free riders* are

peers that eagerly consume resources without reciprocating any in return. It is easy to see how free riders diminish the quality of service for other peers, but more importantly, by making contributing peers feel exploited they disincentivise cooperation in the system and thus threaten the existence of the whole system, especially systems that are predicated on the foundation of sharing.

However, in the context of a decentralised network the most important problem created by free riding is that if only a few users are providing resources, they end up acting as centralised servers, this threatens the security of the network and defeats the very goal of the Web3 application.

Gnutella is a popular peer-to-peer file sharing platform which allows users private access to information. In their paper "Free Riding on Gnutella", Eytan Adar and Bernardo A. Huberman [29] showed that 70% of Gnutella users were not sharing any files and nearly 50% of responses for file discovery were being returned by the top 1% of sharing hosts.

Similarly, Locher at al [30] were able to create "BitThief", a free riding BitTorrent agent that was able to achieve high download rates even without seeding any data in return. They were also able to demonstrate that sharing communities which originally intended to promote cooperation among peers ultimate provide many incentives to cheat.

2) Sybil Attack: In a distributed network, if an entity can control a large number of nodes and hence obtain a large number of node identifiers, they can use this dominance of identities to control the network and undermine the mechanisms of the network which results in a network with less robustness and freedom. Such an attack is often referred to in literature as a *Sybil Attack*, where a *Sybil* is the fake identity of an entity. [31]

The Sybil Attack was first mentioned by Doucer in [32]. In this paper, Doucer argues that only a central authority can prevent a Sybil Attack under realistic assumptions of resource distribution and coordination.

While the intuitive solution to making a network robust against a Sybil Attack seems to be to make it expensive to create new identities in the network, doing so increases the social cost of the network by making it hard for new users to join it.

In the context of reputation mechanisms in decentralised networks, a colluding group of malicious nodes could also increase the reputation of its nodes by itself and hence threaten the integrity of the network

3) Pollution Attack: In 2005, Liang et al [33] showed that it was possible for an attacker in a decentralised network to corrupt certain targeted content, rendering it unusable and then making it available to the network in a large quantity. Since users on the network are unable to distinguish between the polluted and the original content through content discovery alone, users download the polluted content and further share it with other peers, resulting in the polluted content spreading through the network.

In their analysis of the FastTrack peer to peer sharing system, it was found that as many as 50%-80% of copies of popular content were polluted.

4) Index Poisoning: Often resource sharing in decentralised networks is conducted through indices, which allow users to conveniently discover the location of their desired content. Depending on the architecture of the system, the index could be distributed over a fraction of the file sharing nodes (as in FastTrack) or over all the nodes.

In an Index Poisoning attack an attacker inserts bogus records into the index, for example, by inserting random identifiers that do not correspond to any address into the index. This way, when a user attempts to download a file they are unable to locate its content, leading to them finally abandoning the search. [34]

While the Pollution attack described earlier requires the attacker to obtain high-bandwidth to make sufficient versions of the corrupted copies available in the network, the Index Poisoning attack is easier in that it requires less resources to pull off.

5) Slandering: Under Sybil Attack, we discussed that it may be possible for a colluding group of malicious nodes to do *self promotion* to increment their own reputation in a reputation based distributed system. On the other hand, it may also be possible for a group to coordinate to reduce the reputation of a victim, such an attak is called *slandering* [35]

6) White Washing: Nodes that have accrued a bad reputation by acting in an undesired manner can "clean" a bad reputation through *white washing* to avoid the negative effects of the disincentive system [35]

7) Denial of service: Cooperating nodes can work to block the functioning of a distributed systems, preventing other peers from utilizing its services

IV. Evolution of cooperation

One of the foundational works investigating how cooperation can emerge and persist without a third party is "The Evolution of Cooperation", a 1984 book written by political scientist Robert Axelrod which expanded upon the highly influential paper he co-authored with evolutionary biologist W.D. Hamilton [36]. The book's central question is "Under what conditions will cooperation emerge in a world of egoists without central authority?", which seems to be the exact problem Web3 platforms seek an answer for.

Axelrod held two computer simulation tournaments where multiple strategies for playing an iterated two-player Prisoner's Dilemma game were solicited from professionals across multiple disciplines. The Prisoner's Dilemma is a popular game analyzed in game theory where two rational agents are faced with a dilemma, they are arrested by the police and have to individually decide to either cooperate with the police or stay silent. The dilemma was originally framed by Merrill Flood and Melvin Dresher in 1950. A key requirement of the game is that: $t > r > p > s$ and $2 \times r > t$ where t, r, p and s represent payoffs for the different outcomes of the game. If both players choose to stay silent i.e. they cooperate with each other, they are each awarded r, on the other hand if both players defect, they are each awarded s. If one player stays silent while the

other defects, the player who defects is rewarded t while the player who chose to stay silent is paid s. Fig. 2 demonstrates this payoff matrix visually. Hence, although the decision to collectively stay silent is overall the most optimal, individually, the best decision is to defect. Further, in an iterated Prisoner's Dilemma game there is a probability w that two players will interact in the next round. [37] Contestants who submitted algorithms to play the tournament accrued points in each round according to the shown payoff matrix by playing against other strategies. The tournament consisted of five iterated prisoner's dilemma games in total with each game consisting of 200 rounds each.

The Darwinian theory of evolution would suggest that the most selfish strategy would perform the best and while indeed, in a single iteration defecting is always the best strategy, in the iterated Prisoner's Dilemma the strategy that ended up performing the best in both rounds was a simple "Tit For Tat" strategy. As the name suggests, this strategy was based on the concept of direct reciprocity, the strategy's next move is determined by the last move of the opposing strategy, if it cooperated the strategy would cooperate too and conversely, if it defected, the strategy would defect too.

Based on the results of the tournament, Alexrod identified four characteristics that he believed led Tit For Tat to perform the best of all strategies:

1) **Niceness**
   By being nice, Tit For Tat can benefit from long term mutual cooperation with other strategies that are also nice. However, it is important to note that niceness alone would lead to exploitation from other strategies who are not nice
2) **Forgiveness**
   Strategies that are not forgiving are doomed to be locked into mutual destruction after a single defection from an opponent. Tit For Tat allows an opponent to start cooperating again after defecting initially which makes it forgiving
3) **Retaliation**
   As pointed out earlier, niceness alone leads to exploitation by uncooperative strategies. By retaliating when the other strategy doesn't cooperate as expected, Tit For Tat avoids being exploited by such strategies
4) **Certainty**
   By being easy to understand, Tit For Tat makes it easy for other strategies to understand what it's doing thus allowing them to come to a mutually beneficial strategy much faster

Axelrod's analysis thus provides an interesting set of prescriptions for designing strategies for nodes on a decentralised network. Keeping in mind that not all interactions needs to be zero-sum and it may be possible for all cooperating parties to benefit on the long term by cooperating and not being the first to defect seem to work as good principles which suggest that cooperation could indeed organically grow in a pool of egoistic nodes. However, being too nice also has its downsides and any effective strategy should be quick to retaliate to prevent exploitation. Finally, keeping it simple seems to be effective advice otherwise the strategy might risk confusing potentially cooperative neighbours.

Further, there are lessons for designers of Web3 applications, the most important being having a large "shadow of the future", i.e. a sufficiently large w which guarantees that nodes interact with each other more durably and frequently so they have time to develop a mutually cooperative strategy and since they are more likely to defect it the probability of meeting a node again is low. This can be done in many ways including using spatiotemporal structures e.g. clustering of small groups in space [38]

However, there are limitations to Axelrod's results:

1) **Assumptions are too simplified**
   Not all real word interactions are as simple as an Iterated Prisoner's Dilemma game. Often participants can communicate with each other and hence collaboration through other means may be a better strategy. Further, it may not be possible for real world participants to necessarily perceive credible threat, or respond to it rapidly and accurately
2) **Not necessarily universal**
   In his 2000 paper "Twenty Years on: The Evolution of Cooperation Revisited", Hoffman [39] showed that Axelrod's tournament was sensitive to the initial population composition and the potential for strategies to make mistakes. Under different initial compositions and assumptions, other strategies were shown to perform better than Tit For Tat
3) **Does not consider indirect reciprocity**
   While direct reciprocity is a powerful mechanism, it relies on repeated encounters between individuals. However this is too simplifying an assumption to model human interactions where exchanges are often asymmetric and fleeting. Indirect Reciprocity is more representative of real human exchanges where we help people even if they've never directly helped us before based on some indirect exchange. For example, we may donate to a charity which helps other people. [40]

V. Reputation Mechanisms

Instead of only relying on direct reciprocation in distributed systems, we can allow users that help each other out to establish a good reputation which can be used to reward them in some other way. After all, this is more representative of real social interactions, while we are interested in how people interact with us, we are also interested in the actions of others which we learn about from social channels such as gossip. In taking actions, we don't only take into account our direct experiences but also experiences we've learnt about from indirect sources. Similarly, when choosing to assist someone we also consider how it affects our reputation in society.

Although animals possess simple mechanisms for indirect reciprocity, only humans engage in complex reputation systems. [40] This seems to be because such systems require a substantive cognitive load, not only does it require a memory

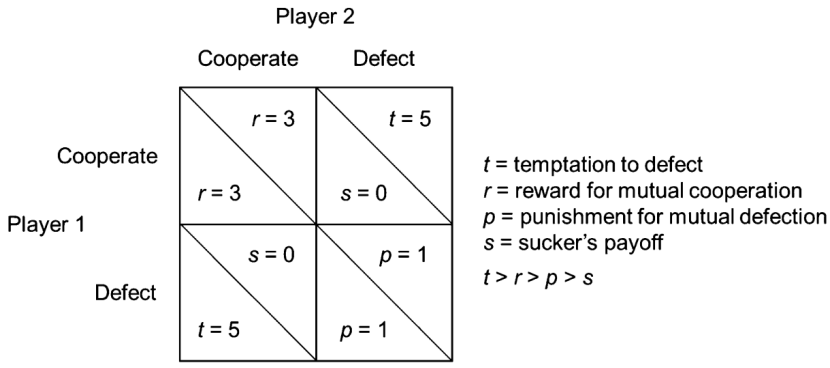

Fig. 2. A typical payoff matrix of a 2 player prisoner's dilemma [38]

of all transactions but also requires the ability to monitor the dynamically changing social network of the group. Hence, the strategies required to succeed in indirect reciprocity are also understandably a lot more complex than the simple Tit For Tat strategy that succeeds in direct reciprocity.

In their paper on reputation systems, Resnick et al [10] define a reputation system as one that "collects, distributed and feedback about participants' past behavior ... these systems help people decide whom to trust, encourage trustworthy behavior, and deter participation by those who are unskilled or dishonest."

As mentioned before, users on decentralised networks look for some form of assurance that their transactions on the network will be successful. The reputation of a user in reputation systems serves as a "shadow of the future" to each transaction, creating an expectation for what a user can expect when dealing with another user.

Consider the example of eBay's reputation system, the "Feedback Forum", after a transaction is completed, a buyer or seller can rate each other (1, 0 or -1) and leave comments. A participant in eBay accumulates such points over time which are displayed next to their screen name. A buyer can view a seller's points and comments left by other users to create a "shadow of the future" into the transaction they can expect to have if they bought an item from the seller. Many other online forums and marketplaces such as Amazon and Stack Overflow rely on similar reputation mechanisms.

According to Resnick, a reputation system must meet three challenges: [41]

1) Provide information that should allow users to distinguish between trustworthy and non trustworthy users,
2) Encourage users to be trustworthy, and
3) Discourage participation from users who aren't

In addition to the above, a successful reputation system should also be able to avoid avoid issues mentioned in III

The following are a few notable reputation mechanisms which attempt to accomplish the objectives stated above:

A. WikiTrust

WikiTrust [42] is the reputation system used for one of the largest collaborative applications known to mankind: the writing of articles on Wikipedia. It is a content-driven reputation system, that is, it relies on automated analysis of the content generated by the user and the collaboration process to derive the reputation of the user, rather than explicit feedback provided by users on other users. It is possible to use such a reputation system since the applications it's catered for is entirely content driven.

The goals of *WikiTrust* are to incentivise lasting, meaningful contributions from users, help increase the quality of content being produced, spot vandals and to offer users an indicator of the quality of the content they are consuming. To achieve these goals, WikiTrust maintains different reputations for users and the content they create.

If a user makes a contribution that is meaningful and its content is preserved in future edits, they gain reputation, on the other hand, if their contributions are wholly or partially undone by future edits, then they lose reputation. Content starts with no reputation, if they are revised by users with high-reputation, it gains reputation. On the other hand, if the text is disturbed by too many edits, indicating that the content may not be trustworthy, it loses reputation.

In order to estimate how much each contribution is preserved or removed as required for the above, WikiTrust relies on an edit distance function $d(r, r')$ which is computed based on how many words have deleted, inserted, replaced and displaced from the edit that led from r to r'. Relying on such a distance functions allows the reputation system to be language independent. Finally, the value of an edit is calculated using the function:

$$q(b|a, c) = \frac{d(a, c) - d(b, c)}{d(a, b)} \tag{1}$$

Where b is the edit being evaluated, a is the revision before the edit and c is the revision after it. $q(b|a, c)$ outputs a value between -1 and +1; it is equal to -1 if $a = c$ and hence implying that b was entirely reverted, on the other

hand, it is equal to +1 if the change from a to b was entirely preserved. However, a limitation of this approach is that since it requires subsequent revisions, it is unable to judge newly created revisions.

WikiTrust only considers not negative reputation values, new users are assigned a reputation very close to 0, this ensures that vandals cannot white wash themselves since their new identities would have a similar reputation to their vandal identity. Also, due to the content driven nature of the system, creating sybils is harder than in a system where identities can simply be used to promote each other.

B. MeritRank

MeritRank [43] uses a merit based tokenomics model which aims to bound the benefits of Sybil attacks instead of preventing them altogether. The system is based on the assumption that peers observe and evaluate each others' contribution, similar to the reputation system used in eBay. Each peer's evaluation is stored in a personal ledger and modelled in a feedback graph where the feedback to each user is modelled as a special token value which accumulates over time. It is also assumed that each peer is able to discover the feedback graph, for example, through a gossip protocol. MeritRank manages to achieve this Sybil tolerance by imposing the following constraints on how reputation can be gained inside the feedback graph:

1) **Relative Feedback**
   This constraint places a bound on how much feedback a single entity can provide to another entity by the degree of the entity i.e. the size of the set of its neighbours. This constraints assists in limiting a single entity from creating multiple parallel sybils
2) **Transitivity α decay**
   This constraint limits the ability of an entity to create a serial sybil attack by terminating random walks in the feedback graph with a probability α
3) **Connectivity β decay**
   Sybil attack edges in a feedback graph are often bridges i.e. their cut creates two separates components. This constraints introduces a punishment for a node for being in a separate component

A trust graph modelled using these MeritRank's constraints will satisfy:

$$\lim_{|S|\to\infty} \frac{w^+(\sigma_s)}{w^-(\sigma_s)} \leq c \tag{2}$$

where, $w^+(\sigma_s)$ is the profit gained by the Sybil Attack σ_s, $w^-(\sigma_s)$ is the cost of the Sybil attack, S is the set of Sybils and c is some constant value such that $c > 0$. Thus MeritRank is able to provide a reputation mechanism with feedback which is Sybil tolerant.

C. PeerReview

TO-DO

D. FullReview

TO-DO

E. ConTrib

TO-DO

VI. Other Mechanisms

Besides direct reciprocity and indirect reciprocity, there are also other mechanisms that should be considered when understanding how cooperation could evolve in a decentralised network. Martin A. Nowak lays out some of these mechanisms in his work "Five Rules for Evolution of Cooperation" [40]:

1) **Network Reciprocity**
   While the analysis so far relies on a well-mixed population, in reality the spatial structures of social connections are not well mixed, instead certain groups interact with each other more often than others. In such a setting, it may be able to form network cluster of cooperators who help each other out resulting in a "Network Reciprocity" which is a generalisation of "Spatial Reciprocity".
   In their paper "The WebEngine - A Fully Integrated, Decentralised Web Search Engine", Mario M. Kubek and Herwik Unger [44] suggest an idea idea of constructing "content overlay networks". This involves creating social graphs with nearby and distant neighbours, where nearby neighbours are neighbours that share similar content.
2) **Group Selection**

3) **Green Beard models**

VII. Conclusion

References

[1] L. Mui, "Computational models of trust and reputation: Agents, evolutionary games, and social networks," 05 2014.

[2] L. Christov-Moore, D. Bolis, J. Kaplan, L. Schilbach, and M. Iacoboni, "Trust in social interaction: From dyads to civilizations," Jun 2022. [Online]. Available: psyarxiv.com/urav8

[3] M. T. F. Zanini and C. P. Migueles, "Trust as an element of informal coordination and its relationship with organizational performance," *EconomiA*, vol. 14, no. 2, pp. 77–87, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1517758013000064

[4] K. Newton, "Trust, social capital, civil society, and democracy," *International Political Science Review*, vol. 22, no. 2, pp. 201–214, 2001. [Online]. Available: https://doi.org/10.1177/0192512101222004

[5] K. S. Cook, M. Levi, and R. Hardin, *Whom can we trust?: How groups, networks, and institutions make trust possible*. Russell Sage Foundation, 2009.

[6] T. Earle, "Trust, confidence, and the 2008 global financial crisis," *Risk analysis : an official publication of the Society for Risk Analysis*, vol. 29, pp. 785–92, 05 2009.

[7] G. Korpal and D. Scott, "Decentralization and web3 technologies," 5 2022. [Online]. Available: https://www.techrxiv.org/articles/preprint/Decentralization_and_web3_technologies/19727734

[8] P. De Filippi, M. Mannan, and W. Reijers, "Blockchain as a confidence machine: The problem of trust challenges of governance," *Technology in Society*, vol. 62, p. 101284, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0160791X20303067

[9] J. Hofmokl, "The internet commons: towards an eclectic theoretical framework," *International Journal of the Commons*, vol. 4, no. 1, pp. 226–250, 2010.

[10] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Commun. ACM*, vol. 43, no. 12, p. 45–48, dec 2000. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/355112.355122

[11] L. Cao, "Decentralized ai: Edge intelligence and smart blockchain, metaverse, web3, and desci," *IEEE Intelligent Systems*, vol. 37, no. 3, pp. 6–19, 2022.

[12] J. Lerner and J. Tirole, "The economics of technology sharing: Open source and beyond," *Journal of Economic Perspectives*, vol. 19, no. 2, pp. 99–120, June 2005. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/0895330054048678

[13] P. Baran, "On distributed communications networks," *IEEE Transactions on Communications Systems*, vol. 12, no. 1, pp. 1–9, 1964.

[14] J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. Epema, M. Reinders, M. van Steen, and H. Sips, "Tribler: a social-based peer-to-peer system," *Concurrency and Computation: Practice and Experience*, vol. 20, pp. 127 – 138, 02 2008.

[15] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Cryptography Mailing list at https://metzdowd.com*, 03 2009.

[16] [Online]. Available: https://www.starlinglab.org/78days/

[17] T. O'Reilly, "What is web 2.0: Design patterns and business models for the next generation of software," *University Library of Munich, Germany, MPRA Paper*, vol. 65, 01 2007.

[18] "Home · solid." [Online]. Available: https://solidproject.org/

[19] G. Wood. [Online]. Available: http://gavwood.com/dappsweb3.html

[20] Z. Liu, Y. Xiang, J. Shi, P. Gao, H. Wang, X. Xiao, B. Wen, Q. Li, and Y.-C. Hu, "Make web3.0 connected," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 5, pp. 2965–2981, 2022.

[21] J. M. Garon, "Legal implications of a ubiquitous metaverse and a web3 future," *SSRN Electron. J.*, 2022.

[22] A. Park, M. Wilson, K. Robson, D. Demetis, and J. Kietzmann, "Interoperability: Our exciting and terrifying web3 future," *Business Horizons*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0007681322001318

[23] J. Bartlett, J. Birdwell, and M. Littler, *The new face of digital populism*. Demos, 2011.

[24] H. T. Williams, J. R. McMurray, T. Kurz, and F. Hugo Lambert, "Network analysis reveals open forums and echo chambers in social media discussions of climate change," *Global Environmental Change*, vol. 32, pp. 126–138, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959378015000369

[25] J. Jiang, X. Ren, and E. Ferrara, "Social media polarization and echo chambers in the context of covid-19: Case study," *JMIRx Med*, vol. 2, no. 3, p. e29570, Aug 2021. [Online]. Available: https://med.jmirx.org/2021/3/e29570

[26] P. Norris and R. Inglehart, *Cultural backlash: Trump, Brexit, and authoritarian populism*. Cambridge University Press, 2019.

[27] J. Schradie, "The trend of class, race, and ethnicity in social media inequality," *Information, Communication & Society*, vol. 15, no. 4, pp. 555–571, 2012. [Online]. Available: https://doi.org/10.1080/1369118X.2012.665939

[28] E. Bellini, Y. Iraqi, and E. Damiani, "Blockchain-based distributed trust and reputation management systems: A survey," *IEEE Access*, vol. 8, pp. 21 127–21 151, 2020.

[29] E. Adar and B. A. Huberman, "Free riding on gnutella," *First Monday*, vol. 5, no. 10, Oct. 2000. [Online]. Available: https://firstmonday.org/ojs/index.php/fm/article/view/792

[30] T. Locher, P. Moore, S. Schmid, and R. Wattenhofer, "Free riding in bittorrent is cheap," in *5th Workshop on Hot Topics in Networks (HotNets)*, 2006. [Online]. Available: http://eprints.cs.univie.ac.at/5696/

[31] J. Dinger and H. Hartenstein, "Defending the sybil attack in p2p networks: taxonomy, challenges, and a proposal for self-registration," in *First International Conference on Availability, Reliability and Security (ARES'06)*, 2006, pp. 8 pp.–763.

[32] J. R. Douceur, "The sybil attack," in *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, ser. IPTPS '01. Berlin, Heidelberg: Springer-Verlag, 2002, p. 251–260.

[33] J. Liang, R. Kumar, Y. Xi, and K. Ross, "Pollution in p2p file sharing systems," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 2, 2005, pp. 1174–1185 vol. 2.

[34] J. Liang, N. Naoumov, and K. W. Ross, "The index poisoning attack in p2p file sharing systems," in *Proceedings IEEE INFOCOM 2006. 25TH IEEE International Conference on Computer Communications*, 2006, pp. 1–12.

[35] V. Agate, A. De Paola, G. Lo Re, and M. Morana, "A simulation framework for evaluating distributed reputation management systems," in *Distributed Computing and Artificial Intelligence, 13th International Conference*, S. Omatu, A. Semalat, G. Bocewicz, P. Sitek, I. E. Nielsen, J. A. García García, and J. Bajo, Eds. Cham: Springer International Publishing, 2016, pp. 247–254.

[36] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[37] R. Axelrod, *The Evolution of Cooperation*, 1984.

[38] J. Barker, "Robert axelrod's (1984) the evolution of cooperation," 01 2017.

[39] R. Hoffmann, "Twenty years on: The evolution of cooperation revisited." *J. Artificial Societies and Social Simulation*, vol. 3, 03 2000.

[40] M. A. Nowak, "Five rules for the evolution of cooperation," *Science*, vol. 314, no. 5805, pp. 1560–1563, 2006. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.1133755

[41] P. Resnick, "Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system," *Advances in Applied Microeconomics*, vol. 11, 10 2002.

[42] L. De Alfaro, A. Kulshreshtha, I. Pye, and B. T. Adler, "Reputation systems for open collaboration," *Commun. ACM*, vol. 54, no. 8, p. 81–87, aug 2011. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/1978542.1978560

[43] B. Nasrulin, G. Ishmaev, and J. Pouwelse, "Meritrank: Sybil tolerant reputation for merit-based tokenomics," 2022. [Online]. Available: https://arxiv.org/abs/2207.09950

[44] M. M. Kubek and H. Unger, "The webengine: A fully integrated, decentralised web search engine," in *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval*, ser. NLPIR 2018. New York, NY, USA: Association for Computing Machinery, 2018, p. 26–31. [Online]. Available: https://doi-org.tudelft.idm.oclc.org/10.1145/3278293.3278294