RTR-TFM: A Routing Threshold-based Randomized Transaction Fee Mechanism

Anonymous Author(s)
Submission Id: 1330

ABSTRACT

In recent years, impossibility proofs have been written claiming the impossibility of achieving efficient and collusion-proof transaction fee mechanisms. In the face of growing consensus that these problems are impossible to solve, this paper offers a dissenting proof, demonstrating the existence of a mechanism that implements the social choice rule of pareto optimality, thereby achieving both incentive compatibility and collusion-resistance.

KEYWORDS

Transaction Fee Mechanism, Leonid Hurwicz, Incentive Compatibility, Free-Riding, Collective Action Problems, Blockchain, Distributed Systems

ACM Reference Format:
Anonymous Author(s). 2024. RTR-TFM: A Routing Threshold-based Randomized Transaction Fee Mechanism. In *ACM Conference, Washington, DC, USA, July 2017*, IFAAMAS, **??** pages.

1 INTRODUCTION

Transaction Fee Mechanisms (TFMs) refer to a class of distributed systems in which a consensus mechanism governs the allocation of the same resource that incentivizes its own provision. Unlike traditional mechanisms, where the number of honest and dishonest processes is static, in TFMs voting power is dynamic — it adjusts with the payouts issued by the mechanism. This introduces the ability for Byzantine strategies that increase profits to compromise the security and stability of the mechanism.

Due to their focus on the technical properties of systems, computer scientists often name these attacks after the "mechanism-specific" techniques they exploit, resulting in a wide array of terminology such as sybil attacks, block-orphaning attacks, selfish mining attacks, fee manipulation attacks, eclipse attacks, side-contract payments, and others. While most researchers treat these vulnerabilities as isolated technical challenges, a few scholars have applied concepts from mechanism design to ask whether general solutions are theoretically feasible. Unfortunately, this line of research has led to a series of impossibility results, suggesting that socially optimal TFMs may be infeasible.

This paper challenges these impossibility results by identifying the specific economic equilibrium in which all such attacks become irrational. We argue that three distinct types of goal conflict -— self-interest, free-riding, and strategic manipulation —- are the key factors preventing this equilibrium from being realized in most *TFMs*. We then review several commonly cited papers and demonstrate that their conclusions stem from their reliance on auction models rather than market models, a choice which limits their ability to address all three types of goal conflict or handle the informational complexity necessary to compute the required equilibria.

Using the language of mechanism design, this paper demonstrates that the social choice rule needed to achieve fee-optimality and collusion-resilience is *pareto optimality*, but the direct mechanisms used to model TFMs are incapable of implementing this rule, as doing so requires multi-dimensional preference revelation across a high-dimensional preference space – a level of informational complexity that composable algorithms cannot handle. While Maskin's Revelation Principle suggests that a direct mechanism must exist for any indirect mechanism, in this case achieving optimality requires decomposable algorithms that use the "no-trade option" to reduce the complexity of computation and limit the scope of the state transitions proposed to those consistent with an efficiency shift towards *paretop optimality*.

Since familiarity with economics is needed to understand why these problems exist, the next section of this paper begins by identifying the novel characteristics of TFMs, and showing how they create three distinct kinds of goal conflict. We then show why *pareto optimality* is the social choice rule needed to eliminate all three, which leads to a review of the impossibility results mentioned above and a discussion of how the conclusions of these papers reflect the informational limitations of the auction models they use to analyse the problem.

In the second half of this paper, we then introduce a novel class of indirect mechanism that is theoretically capable of achieving *pareto optimality*. We provide the formula for this mechanism and show that it's behavior is inconsistent with the impossibility proofs offered in the general literature and that it successfully manages to eliminate all three types of goal conflict. We then provide a game-theoretic treatment of the mechanism to provide a formal proof of incompatibility with the impossibility results this paper debunks.

2 THE NOVEL CHARACTERISTICS OF TRANSACTION FEE MECHANISMS

The novel characteristics of *TFMs* that lead to suboptimal provision are *non-excludability*, *self-provision* and *informational decentralization.*

Non-Excludability allows anyone to use or provision the networks on equal terms provided they are willing to pay the competitive market price. This economic characteristic underpins the technical properties of *censorship resistance*, *decentralization* and *network*

ACM Conference, , July 2017, Washington, DC, USA. © 2024 Association for Computing Machinery. ...$ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
...$15.00

resilience: censorship requires a mechanism with the power to exclude; centralization creates barriers to entry; resilience comes from the ability to route around byzantine actors by adding new nodes to the network. Non-excludability also contributes to economic efficiency in *TFMs*, as efficiency is maximized when producers build atop blocks proposed by their peers rather than orphaning them.

Informational Decentralization refers not to the casual concept of *decentralization* as used in computer science (see: non-excludability) but the economic definition offered by Hurwicz (1972) for mechanisms in which "participants have direct information only about themselves." This characteristic makes *TFMs* vulnerable to Byzantine strategies, as identified by Hurwicz, in which participants manipulate the informational environment that others rely on to make strategic decisions.

Self-Provision allow *TFMs* to support themselves without an owner, relying instead on payouts to network participants. While volunteer-run networks are theoretically possible, their designs fall outside the scope of *TFMs* as transaction fees are purely redistributive. For this reason, in volunteer mechanisms the imposition of fees leads to a dead-weight efficiency loss, since any fee-level above zero is strictly suboptimal given the cost structure of the network.

These three characteristics create fundamental tensions that *TFMs* struggle to reconcile. They must permit open access without enabling sybil attacks, offer private benefits without socializing losses, and use decomposable algorithms while resisting byzantine manipulation of the information environment. We can see the importance of all three characteristics from the way they form an *economic trilemma* where the removal of any one property offers immediate relief to the problems created by the other two.

Understanding these characteristics allows us to identify the three types of *goal conflict* that drive byzantine attacks on *TFMs*. The first type, conflict rooted in *self-interest*, occurs when participants prefer to allocate their resources differently than the mechanism designer intends. For example, a user might desire to save a portion of their transaction fee to spend on other goods and services, rather than adhering to the mechanism's optimal allocation. In this case, participants are signalling disagreement with the designer's intended allocation of utility, both *within* the mechanism and *between* the mechanism and other external goods. These attacks consequently involve participants choosing to bid at suboptimal fee levels, as they prioritize their personal preferences over the collective optimal outcome.

Our second form of goal-conflict is *free-riding*, which emerges because the combination of *non-excludability* and *self-provision* creates public goods within the consensus mechanism. While free-rider pressures are common in many mechanisms, in *TFMs* they are more intractable due to the presence of a dual-sided free-rider problem where users and producers can free-ride on the mechanism in different ways: producers by maximizing the revenue they extract from any collective payout like the block reward, and users by minimizing their contribution to the security budget. As our next section explains, these are the class of attacks that manifest in the form of side-contract payments.

Our third form of goal-conflict is *strategic manipulation*, which emerges because – as Leonid Hurwicz observed – in informationally-decentralized mechanisms participants can strategically manipulate others into suboptimally allocating their own resources by manipulating the informational space in which they form their own strategies. This class of goal-conflict incentivizes producers to create fake transactions, and users to exploit threshold vulnerabilities in auction designs. It is the main problem mechanism designers eliminate when they design mechanisms that achieve bayesian incentive compatibility or incentivize the truthful revelation of preferences.

While conflict over self-interest, free-riding and strategic manipulation are all distinct types of goal-conflict, each type has different causes and manifests in different ways. This is the reason incentive compatibility seems so intractable in *TFMs*, as techniques intended to prevent *strategic misrepresentation* cannot eliminate goal conflict entirely unless conflicts over self-interest and free-riding are also addressed. Any general solution requires the *TFM* to implement an equilibrium in which none of these conflicts exist, which is why the next section pulls back to economic theory to show why *pareto optimality* must be the social choice rule chosen by mechanism that seeks optimal fee-throughput in a collusion-free equilibrium.

3 PARETO OPTIMALITY AND ITS VULNERABILITIES

In the field of economics, the pioneering work on welfare optimality was the publication of Vilfredo Pareto's "Cours d'économie politique" (1896), which introduced the concept of pareto optimality. Pareto defined this state as one where resources are allocated so efficiently that it is impossible to improve overall social welfare by changing the way in which resources are allocated to the production of utility.

From a mathematical perspective, Pareto optimality is achieved when the marginal utility derived from the last unit of each good purchased by each individual is proportional to its production cost. This implies that individuals are spending their resources in a way that maximizes their utility -— essentially, every dollar is spent on whatever good or service provides the greatest marginal benefit to its consumer. This allocation is considered individually rational and provides two important social critera demanded by *TFMs*. First, mechanisms in a *pareto optimality* equilibrium are free from conflicts involving self-interest since no party will unilaterally desire to pay a greater or lesser fee. Second, *pareto optimality* has attractive collusion-proof properties: if no individual can reallocate his own resources without making himself worse off, no group of similar individuals can collude to do so without at least one member of the group suffering as a result. This eliminates all categories of user-user and producer-producer collusion.

But is it possible for *pareto optimal* equillibria to be robust against goal-conflict involving *free riding* pressures or *strategic manipulation*?

The first question was addressed by Samuelson (1954) when he observed that achieving optimal production levels is challenging for goods with non-excludable benefits. If users can lie about the utility they receive from such goods they can pay a lower fee themselves while enjoying the higher level of utility funded by contributions from their honest peers. It was Samuelson's demonstration of this problem – that individual rationality subverted pareto optimality – that led **?**] to coin the term *incentive compatibility* in reference to the opposite condition, the state in which the utility-maximizing behavior of individuals is *compatible with* or leads emergently to the desired welfare condition referred to as *social choice rule.*

Samuelson's observation is why free-rider pressures constitute the second type of *goal conflict* within *TFMs*, where they manifest in the form of side-contract payments. From the perspective of users, selling transactions to block producers gives producers the ability to collect their fees without the need to compete so intently for the privilege. Producers will happily accept a lower fee from users as less of their own income need flow into the public security budget. This form of collusion involves producers helping users free-ride on the contributions of other users to the collective security budget, as analogous to the classic free-rider in Samuelson's model.

On the producer side, side-contract payments permit block producers to free-ride on their peers as well. In this second case, producers offer users transaction-inclusion at suboptimal rates because private control of the transaction fee expands the producer's share of blocks committed to the longest-chain, allowing them to extract more income from any non-excludable payout like the block reward than they lose by subsidizing the user's transaction. Once again we have a situation analogous to Samuelson's model, except in this case the incentive to collude comes from producers and the incentive is to collect more in revenue not pay less in fees.

Understanding the two-sided nature of free-riding in *TFMs* is critical for designing mechanisms that eliminate this form of goal conflict. In the absence of this understanding, it is common to consider all forms of user-producer agreement as suboptimal. But this is not the case! If price negotiations between users and producers drive the cost of blockspace towards *pareto optimal* levels without affecting the overall level of public good provision, they technically shift the network into a more efficient equilibrium in which fee-throughput level are more optimal and *goal conflict* is avoided. It is also trivial to see that side-contract payments will never drive transaction fees below the cost of blockspace in the absence of public goods, as rational producers cannot sustainably accept transaction fees that are lower than their private cost of providing blockspace.

The inability of proof-of-work and proof-of-stake designs to contain fundamental pressures to free-ride is a major cause of inefficiency and suboptimality in those networks. As we shall see, these pressures are also responsible for a non-trivial number of impossibility results, since the techniques mechanism designers use to prevent other classes of goal conflict – such as inducing truthful preference revelation – can contain adversarial forms of strategic manipulation but fail to prevent the sorts of cooperative attacks we see with free-riding strategies.

Our first two classes of goal conflict are thus "self-interest" and "free-rider pressures". The first exists in mechanisms that lack *pareto optimality* and can be solved only by designing mechanisms that implement that social choice rule. The second subverts the ability of mechanisms to achieve *pareto optimality* and can only be rectified by eliminating the public goods that lurk within their incentive sub-structures.

This leaves our third category of *goal conflict*, which is the practice of *strategic manipulation.* To put this issue in historical context, it is useful to know that by the late 1950s and 1960s, the problems that Samuelson flagged regarding the efficient provision of public goods had become widely accepted in mainstream economics. Nonetheless, most economists still believed the production and trade of private goods under classical assumptions was more-or-less *pareto optimal.* Or so they believed until 1972 when **?**], in his second great contribution to mechanism design, pointed out that similar problems also subvert the *pareto optimal* provision of private goods in informationally decentralized mechanisms.

The cause of the suboptimality Hurwicz identified came from the need for participants to exchange information as part of their price-discovery process. In any situation where agents could manipulate the informational environment they could theoretically induce others to strategically misallocate their own resources. The particular passage in Hurwicz's paper that points this out is worth quoting in full:

> Economists have long been alerted to this issue by Samuelson (1954) in the context of the allocation problem for public goods. But, in fact, a similar problem arises in a "nonatomistic" world of pure exchange of exclusively private goods.... If [two parties] were both told to behave as price-takers it would pay one of them to violate this rule if he could get away with it. Now we assume that he cannot violate the rule openly, but he can "pretend" to have preferences different from his true ones. The question is whether he could think up for himself a false (but convex and monotone) preference map which would be more advantageous for him than his true one, assuming that he will follow the rules of price-taking according to the false map while the other trader plays the game honestly. It is easily shown that the answer is in the affirmative. Thus, in such a situation, the rules of perfect competition are not incentive-compatible.

In this case, our form of goal conflict does not involve participants re-allocating their own resources (self-interest) or cooperating with others to exploit public goods (free-riding) but adversarially manipulating the informational environment to frustrate efficient price-discovery. In the context of *TFMs*, we see this exploited whenever producers put their own fees into blocks, or costlessly loop money around the chain.

Awareness of these informational attacks is what led Hurwicz to develop his framework for studying *incentive compatibility*, which asks whether specific market structures (mechanisms) can achieve (implement) specific outcomes (social choice rules) in the presence of participants who make strategic decisions on the basis of private information. This is the reason "truthful revelation of preference" is considered such an attractive property in mechanism design, as it implies the mechanism is not vulnerable to this particularly category of goal conflict.

As an aside, since several papers on *TFMs* declare *incentive compatibility* impossible to achieve, it is useful to remember that Hurwicz never made this claim. As his student Eric Maskin later pointed out, such claims show a misunderstanding of the framework, since all mechanisms are by definition incentive compatible with their outcomes. What a failure of incentive compatibility means is that if private preferences are used to form the strategies adopted by participants in a mechanism, then without an "incentive for truthfulness" mechanisms cannot be assumed capable of implementing any social choice rule.

This point is important for ultimately implementing pareto optimality within a distributed system. For while Hurwicz is often misinterpreted as implying that the direct revelation of preferences is a pre-condition for achieving *pareto optimality*, the truth is more nuanced – market structures still exist which lack the problems Hurwicz identified with *strategic manipulation*, the key exceptions being *atomistic* markets characterized by perfect competition, markets in which the utility purchased varies with price paid, and markets lacking a pre-exchange messaging step. Eric Maskin, who later win the Nobel Prize for his work on the revelation principle, confirmed Hurwicz's intuition when he found that *pareto optimality* is possible in some market structures without the need for truthful preference revelation as an intermediary step.**?**]. His revelation principle also illustrates this in a more subtle way, by showing that a symmetry of outcomes must exist between mechanisms where information is computed in decomposable fashion using agent-level functions, and mechanisms where the exact same information is revealed truthfully and the computation is performed by a centralized mechanism in a non-decomposable fashion. As Maskin showed, if the centrally-computed outcome does not result in a Nash Equilibrium then the decomposable function cannot have one and at least one agent must be lying about their true preferences.

Maskin's work revealed a deeper truth: all incentive compatible mechanisms will induce the revelation of private information one way or the other, meaning that the difference between mechanisms is not whether they reveal user preferences so much as whether they reveal those preferences *directly* or *indirectly*. In direct mechanisms participants share their preferences truthfully in the pre-exchange negotiation step, while in *indirect* mechanisms they reveal them either obliquely in the price-discovery process (such as by negotiating for bundles of goods) or by skipping the price-discovery stage and simply submitting purchase orders directly onto the market.

Back on topic, since *TFMs* are *informationally decentralized* mechanisms that involve users and producers making strategic decisions on the basis different preferences for the allocation of resources within the mechanism, if our social choice rule is *pareto optimality*, we cannot achieve it in any mechanism where participants can costlessly mislead others by manipulating any information relevant to fee-levels in the state of consensus. If a mechanism permits block producers to costlessly include their own transactions in blocks we thus have *de facto* grounds for concluding that incentive compatibility with *pareto optimal* fee-throughput will be impossible to achieve in that mechanism. Strategic manipulation can only be eliminated in mechanisms that make the inclusion of self-generated transactions costly, such that the decision by a block producer or user to include their own fees in a block reveals private information that the mechanism can exploit to shift its own provision into a more efficient equilibrium.

Hurwicz (1973) provides several other conditions any *TFM* will need to meet in order to successfully implement *pareto optimality*. The first is that one-shot mechanisms are insufficient, since algorithms with *inertia* are required to iterate price levels into their optimal positions over time [CITE]. This suggests that the information required to calculate the price of blockspace must be somehow calculable from the state of consensus rather than collected exclusively from peers. And as Jordan (1986) observes, some form of smoothing of costs or payouts is beneficial to prevent mechanisms from unpredictably oscillating around the desired equilibrium point. As we shall see in the next section, these requirements are also incompatible with the vast majority of papers attempting to model the feasibility of building a dream TFM.

In summary, our three types of goal conflict – self-interest, free-riding, and strategic manipulation – are distinct issues that affect most *TFMs*. All three undermine the ability of any mechanism to achieve *pareto optimality* which in turn prevents them from targeting a highly efficient and collusion-proof equilibrium. Each type of goal conflict manifests as unique technical attacks involving different actors, different types of messaging, and targeting different steps in the operation of the consensus mechanism. A block producer who floods the network with spam transactions to drive up fees is engaging in strategic manipulation. A threshold user who underbids in a Vickrey-Clarke-Groves auction is exhibiting self-interest. Users who conspire with producers to defund the security budget are free-riding on their non-colluding counterparts. All three classes must be eliminated to achieve an optimal *TFM*, which is why is why achieving it is so difficult in practice.

With this framework in place for understanding the categories of problems *TFMs* face, in the following section we turn our attention to the existing literature on *TFMs* in computer science, with the goal of showing why the impossibility results in these papers reflect the limitations of their models rather than the limits of what is possible in distributed systems.

4 THE TRANSACTION FEE MECHANISM LITERATURE IN COMPUTER SCIENCE

To our knowledge, this is the first paper to show how goal conflict prevents *TFMs* from achieving *pareto optimality* and makes auction models informationally incapable of resolving conflict within *TFM*. In order to understand the exact problem, this section reviews how computer scientists have studied this issue to show the general issues with approaches used.

Early attempts to model *TFMs* as auctions were Bitcoin-specific, starting with "Redesigning Bitcoin's Fee Market", which proposed using a "monopolistic auction" to stabilize miner revenue, followed by Andrew Yao's "An Incentive Analysis" which showed this maximized miner revenue at scale. Basu, Easley, O'Hara, Sirer then proposed a modified Vickrey-Clarke-Groves mechanism as a better choice for maximizing the collective welfare of both users and miners.

While all three papers focused explicitly on Bitcoin, the concerns over efficiency showed awareness *TFMs* are not just resource allocation mechanisms, but are themselves subject to conflict over resource allocation within the broader economy! Computer science was on the cusp of seeing the underlying economic nature of their problem, identified by Hurwicz in 1973 as "goal conflict", and realizing that *pareto optimality* would be the social choice rule required to solve it.

Computer science pulled back slightly in 2021 when Tim Roughgarden**? ?**] offered a paper that modelled Transaction Fee Mechanisms (TFMs) as two-sided auctions in which block producers are given a temporary monopoly over the production of a block and must strategically allocate a subset of transactions into it. Looking beyond Bitcoin towards a landscape of competing consensus mechanisms, Roughgarden returned to characterizing the incentive-alignment issue as resulting from internal rather than economy-wide conflict over resource allocation. He was the first to highlight the difficulty of achieving incentive compatibility for both users (UIC) and block producers or miners (MIC), leading to seminal works [**? ?**] on the limitations of Bitcoin's "first-price auction" and Ethereum's EIP-1559 [**?**] among others. Roughgarden [**? ?**].

Since 2021, the vast majority of academics working on *TFM design* have followed Roughgarden in modelling *TFMs* as two-party auctions in which producers clash with users over how to allocate blockspace. The attractiveness of the approach is obvious: it focuses on internal rather than external motivations for conflict, it targets an essential step in the formation of consensus, and it uses a two-sided game that is tractable to model. Significantly, Myerson's lemma and virtual valuations can also be used to generalize the rational strategies of participants in these games so they can be asserted to hold in larger games with many players. Unfortunately, the work is simply producing a series of impossibility results.

Since one of the purposes of this paper is to present a mechanism that evades these problems, it is useful show how this choice of modeling *fee mechanisms* has created structural incompatible with a productive solution. In this light, the first problem is methodological treatment of UIC and MIC as properties which can exist outside the context of a social choice rule. Instead of identifying an equilibrium like pareto optimality that guarantee both fee-optimality and collusion-resistance for users and producers alike, and asking what private information both participants would need to disclose for any *direct mechanism* like an auction to achieve it, the literature assumes that truthful preference revelation is a sufficient goal in-and-of-itself. This is typically done by citing the Revelation Principle and observing that any mechanism capable of achieving a nash equilibrium must have an equivalent in which (see Roughgarden p. 13) truthful bidding is a dominant strategy.

The problem with this assumption is that the preference information any algorithm needs to be revealed depends on the social choice rule at stake, and specifically on whether we are in the presence of a problem that requires high-dimensional preferences to calculate.

Viewed sympathetically, we can intuit that the field's implicit social choice rule is an "efficient allocation" of blockspace. This seems fair to assume given Roughgarden's citation of the Vickrey-Clarke-Groves (VCG) mechanism as being UIC and the lack of any seeming challenge to this assumption. If this auction is considered to reveal truthful information sufficient for optimizing participant utility in one mechanism, it seems intuitive that it would collect the same information needed to optimize utility in a different mechanism, but the information required actually depends on the social choice rule and the difference between the types of conflict mechanisms are intended to address requires a very different type of "utility" information to be collected in our case.

Note, for instance, that the VCG auction is a *direct mechanism* that does not require high-dimensional preference information as part of its process of truthful preference revelation. Users share information on the maximum price-point at which they are willing to purchase the single privte good being allocated given a fixed price and production schedule for everything else, not their comparative preference for how to divide their resources between all goods and services competing for consumption of the same transaction fee at all viable price equilibria as required for implementing *pareto optimality*. The VCG auction is thus informationally inadequate for eliminating byzantine strategies motivated by "self-interest" – our first class of incentive to suboptimality. Similarly, the VCG auction has no informational basis for combatting *free-riding*, since those are cooperative strategies to defund the production of a form of utility not covered by models that treat blockspace like a private good.

Roughgarden's papers were quickly followed by papers from Elaine Shi and Hao Chung, who offered technical definitions like "side-contract proof" ("no utility increase from off-chain payments") rather than using Roughgarden's technical definition of OCA-Proof. The difference between the two is essentially the difference between whether collusion maximizes revenue in the context of a single block or across potential forks in a chain. The concept of OCA-Proof thus encompasses types of collusion that lead to block orphaning while SCP-based approaches do not. Pareto optimality eliminates both possibilities on the fundamental grounds that there is no rational strategy for colluding in either case. The mechanism proposed later in this paper also elegantly sidesteps Roughgarden's concerns that any "fee burn" must invite collusion because "because OCAs allow the miner and users to coordinate and evade the intended burn." This is of course not possible in routing work mechanisms where the burn is the cost of producing a block, as it cannot be evaded by moving the payment off-chain.

A more subtle problem applies to the treatment of block producers, who are simply asked to implement the fee mechanism. The lack of any need for producers to reveal private information raises questions about why we are modelling this game as a two-sided strategic interaction, and points to a deeper methodological problem. For as noted in our first section, the class of *TFMs* we are studying contain dual-sided free-rider problems. This specific class of vulnerability makes it impossible to achieve pareto optimality for both parties if we require truthful preference revelation from only one party. For both parties have private incentives to adopt byzantine strategies that are driven by a desire to free-ride on their peers. Eliminating collusion thus requires either eliminating collective action problems generally (and the auction mechanism cannot handle this as it focuses exclusively on a single private good) or by identifying a kind of "private information" which can by leveraged by a mechanism to motivate producers to shift their strategies away from defunding fection and towards cooperation. By denying producers the ability to act strategically on the basis of private preferences, modelling blockchains as auction mechanisms leads inescapably to impossibility results as they prevent the most critical party from behaving strategically!

The third and most fundamental problem with the auction model generally is that it is impossible to generalize its impossibility results, since the existence of an impossibility proof for this specific type of *direct mechanism* can never eliminate the possibility that an *indirect mechanisms* might exist that can achieve the desired results through the solicitation of a different kind of preference. Since this is a somewhat subtle point, note that while Maskin's relevation principle teaches us that all nash equilibria which are reachable by *indirect mechanisms* can be implemented as *direct mechanisms*, the opposite is not true. So even if the auction model was informationally appropriate for implementing *pareto optimality*, we cannot conclude from an impossibility proof generated assuming the limitations of a direct mechanism that no indirect mechanism exists which is capable of skirting that problem.

Understanding this point is important for seeing how the mechanism described later in this paper solves the problem. For Maskin's revelation principle is based on logical reasoning about the consistency of outcomes between decomposable algorithms (where participants compute their preferences privately) and composable algorithms (where users reveal their preferences to a centralized mechanism that does the work for them). In situations where the amount of information required to calculate an optimal solution is so large as to make disclosure impractical or impossible to calculate in a centralized mechanism, such as exists with the high-dimensional preference data needed to compute *pareto optimal* equilibria in informationally decentralized environments, *indirect mechanisms* that use *decomposable algorithms* to filter and transform participant preferences prior to their revelation can be informationally necessary to achieve incentive compatibility. It should be noted that Maskin's revelation principle still holds – truthful preference revelation happens in both types of mechanism – but it can happen in a different stage, either in the "action stage" identified by Hurwicz where bids are submitted directly to the market, or obliquely in the "pre-exchange negotiation stage" in a more indirect and filtered form.

The presence of public goods in consensus mechanisms is what forces the need for high-dimensional preference measurement, as they pull focus away from maximizing the kind of "single well-defined objective function" Hurwicz associated with computational models and towards the more complicated multi-variate forms of goal conflict suitable for economic analysis. Perhaps because of this, it is not surprisingly to see mechanism designers with stronger economic backgrounds explicitly recognize the presence of public goods, as is the case in a recent paper by Elijah Fox, Mallesh Pai, and Max Resnick on "Censorship Resistance in On-Chain Auctions". While the assumptions these authors make are not strictly true – transactions fees induce both private and public goods and only incentivize the provision of public goods to the extent they circulate openly for competitive inclusion – these authors are absolutely correct that off-chain payments involve a form of free-riding and addressing this problem is the key challenge for mechanism designers.

While the proposal by Fox fails on technical grounds (the degree of competition for fee collection can be manipulated by collusion in any non-excludable mechanism), their insight helps explain why *indirect mechanisms* are traditionally used in economics when optimizing resource allocation to non-excludable goods and eliminating free-rider pressures. *Indirect mechanisms* are the preferred approach for solving these class of problems, such as in the curious case of the Clarke-Groves mechanism (not to be confused with the VCG mechanism), a indirect mechanism in which users are asked to submit bids across bundles of goods, some of which may include public goods. Given the parallels between the information requirements to solve both problems, it is likely no accident that the solution this paper identifies is an *indirect mechanism* that leverages decomposabitility to avoid the need for truthful preference revelation during the "pre-exchange negotiation step" as a necessary precondition for achieving incentive compatibility.

Returning to our review of the related literature, a second influential string of papers has come from Hao Chung and Elaine Shi in their work on *side-contract payments* and the *zero-revenue bound.* Specifically, Hao and Chung advance claims that side-contract payments (SCP) are impossible to disincentivize in any mechanism where the income for block producers is above zero.

The same methodological problems apparent elsewhere replicate here, as Hao and Chung treat truthful preference revelation as if it is a valid social choice rule rather than an intermediary step to achieve one in the presence of private information. The most significant difference with this approach is that unlike academics who view an "efficient and fee-stable blockspace allocation" as the implicit social choice rule, for Hao and Chung it is the possibility for a collusion-free environment that takes center stage, with results suggesting that collusion is impossible to eradicate in *TFMs* with revenue above the *zero-revenue bound.*

The framework provided above provides an intuitive explanation of why Hao and Chung stumble into their zero-revenue bound. As explained in Section 2, the underlying source of suboptimal forms of user-producer collusion is the existence of dual-sided free-rider problems embedded in the mechanisms. Their results follow deductively from this problem. At any positive fee-level users have an incentive to collude with producers to free-ride on the contributions of their peers to the security budget. This problem can be avoided by compensating producers through an inflationary block reward, but that invites producers to free-ride on the supply-side payout. Avoiding one trap pushes us into the other, so the only situation in which we avoid collusion completely in their model is if neither fees nor block rewards exist.

What percentage of the remaining papers are writing about blockchains and what percentage are simply writing about auctions? Making similar assumptions as their predecessors (auction model, no clear social choice rule, costless manipulation of informational environment), Aadityan Ganesh, Clayton Thomas and Matthew Wienberg not surprisingly end up in the same place, with the value of their work consisting mostly of several new terms like "off-chain influence proofness" that capture specific forms of collusion. While the authors identify "external opportunities" for profit not captured within the fee mechanism (implying goal conflict), they fail to follow their observations to their obvious conclusions: that the auction model itself is an inappropriate tool for analysing this problem. But the proof-of-stake models they study cannot address any of these problems. So how could they – or any of their peers – be expected to find a solution, when their focus is examining mechanisms designed by developers who also fail to understand the underlying problems they face?

There are some positive results, interesting primarily for showing that market mechanisms – not auctions – hold the key to solving these problems. Rejecting the tendency to treat auctions as one-shot games,CITATION find that repeated-games. This works for the same reason that **?**] finds – it creates the form of "inertia" required in free markets for price levels to move closer to pareto optimal levels in equilibrium. Ferreria's finding also shows the benefit of moving pricing information into the state of consensus itself, minimizing opportunities for *strategic manipulation* by shifting the market price into the environment rather than making it only accessible through unreliable peer messages.

The tendency in computer science papers to treat the conclusions of earlier papers as axioms in new lemmas intended to develop new theorems has exacerbated the tendency for impossibility results to be exaggerated and amplified.

?] introduce a "Burning Second-price" TFM that compromises allocative efficiency to guarantee user and block producer IC. In their model, the authors tweak the utility model with "γ-Strict" utility for users/producers. The new model captures the future cost of introducing fake transactions discounted by a *public* parameter $\gamma \in [0, 1]$. We believe compared to "γ-Strict" utility RTR-TFM's incentive rule introduces a natural cost for introducing fake transactions to the users/producers. Moreover γ-Strict utility does not prevent free-riding.

5 RTR-TFM: A PARETO OPTIMAL SOLUTION FOR DISTRIBUTED CONSENSUS

As can be seen from Appendix A, the mechanism that eliminates all three categories of goal conflict is an *indirect mechanism* that implements *pareto optimality* as its social choice rule. The mechanism is reverse clock auction where producers compete for the right to produce blocks by collecting transactions and burning their fees. The efficiency with which they perform this task affects how quickly they can produce the next block, and also affects the probability that they will be selected after-the-fact to receive a partial payment as one of the many routing nodes that contributed to the production of the block.

The mechanism leverages the fact that including topological information in blocks creates an efficiency gradiant that allows the consensus-layer to asymmetrically punish attackers without the need to identify them. The reason for this is that as long as a subset of honest nodes are capable of proposing at least a subset of blocks more efficiently than an attacker (i.e have a more compact routing path for a subset of network transactions), attackers will need to shift the network into a more inefficient topology in order to collect the same fees, since the attackers will be at least one hop deeper in the network for at least one transaction. This shift is informationally visible to consensus through the existence of cryptographically-secured routing signatures embedded in transactions, which allows the mechanism to increase the cost of block production as well as reducing the probability of the attacker collecting the payout relative to the honest node whose block is being orphaned or censored.

The mechanism induces both users and producers to the truthful revelation of preferences. Users reveal the utility they get from inclusion in the mechanism. Producers reveal their comparative efficiency at extracting profits from the transactions in their mempool. Nodes that are highly-efficient can spend their own money to produce blocks, this increases their chance of contributing the next block at the cost of lowering the marginal profitability.

The mechanism thus functions like a market. Users compete with users for transaction inclusion. Producers compete with producers for profits. The mechanism is competitive on both sides, and thus has the properties of a "greed process" a mechanism that is proven to be incentive compatible with pareto optimality in informationally decentralized environments without the need for truthful preference revelation as an intermediary step as in *direct mechanisms* like auctions..

In the section that follows, we provide a game theoretic proof of these claims, demonstrating:

* it is costly to use one's own money to produce blocks ceteris paribus * block producers may spend fees to speed * in this situation, the producer accepts lower profits

Users allocate their resources to the blockchain based on the utility that. Producers compete to provide the network and collect fees by proving their efficiency at doing so. The mechanism , turnin

RTR-TFM: Routing-based Randomized TFM

In this section we introduce RTR-TFM: a Routing Threshold-based Randomized-TFM. Figure **??** presents the overview of our mechanism, which is described briefly below.

There are three key differences between RTR-TFMand existing TFMs. **First,** we introduce the concept of "routing-work" which replaces *mining* or *staking* as the form of work used to regulate the pace of block production. The amount of routing work in any block is verifiable in an informationally decentralized and permissionless environment and provides an objective standard for determining when a block can be produced. **Second,** we propose a payment rule that makes payments available to multiple network nodes and pulls payouts away from nodes that orphan blocks with statistical regularity. **Third**, we demonstrate how competition between users and block producers in mechanism eliminates opportunities for the socially-suboptimal forms of value-extraction that prevent attaining pareto optimality in other mechanisms. Figure **??** summarizes the main components of RTR-TFM.

In RTR-TFM, when users send transactions to nodes in the network, they include cryptographic routing signatures indicating the first *hop* node. Each node adds its signature as it *propagates* the transaction deeper into the network, creating within each transaction an unforgeable record of the path the transaction has taken from the user to the block producer offering inclusion.

The "routing work" that nodes need to produce blocks and which regulates payouts is derived from this chain of signatures. Specifically, the amount of routing work that is available to a potential block producer from any transaction is given by $c \cdot \frac{1}{2^{h-1}}$, where c is a network-determined constant and h is the node's hop for that transaction. E.g., a node hearing about a transaction at its third hop receives $\frac{c}{4}$ routing work for that transaction. Each node gathers transactions until they have enough total routing work to meet a network-determined *difficulty threshold*, τ. At this time, the node may become a block producer and broadcast its block with its set of transactions whose total routing work crosses τ.

Once a block is produced, we *burn* half of the total fees. The other half is given to a *random* routing node selected from the routing paths within the transactions in the block, selected using the following probability distribution: (i) we uniformly select a transaction among those that are part of the block weighted by the total fees paid by that transaction, then (ii) from the routing path of the winning transaction, we randomly select a routing node, with each node weighted by the routing work available at their specific hop to the total amount of routing work generated across the entire relay path.

Importantly, while the block producer is eligible for payment as the final routing nodes in the routing paths of every transaction, other nodes also have a probability of being selected for payment. This is an important distinction between RTR-TFMand other mechanisms, as block producers who include their own fees in blocks are not guaranteed to recapture them.

5.1 Game Theoretic Characterization

TFM Model [? ?]. We model RTR-TFM as a game with a set of $m \in \mathbb{N}$ block producers, $\mathcal{P} := [m]$. We consider each block producer $i \in \mathcal{P}$ to be *myopic* and *strategic*. We assume that each transaction is of the same size, with each block's capacity denoted by $k \in \mathbb{N}$. Furthermore, let $n \in \mathbb{N}$ denote the total number of users, denoted by $\mathcal{U} := [n]$. We assume that each user $j \in \mathcal{U}$ is also myopic and strategic [? ? ? ? ?]. A user $j \in \mathcal{U}$ is interested in getting a slot in the block for its transaction. Let $\theta_j \in \mathbb{R}_{\geq 0}$ denote user j's private valuation for its transaction's confirmation and $b_j \in \mathbb{R}_{\geq 0}$ as its transaction's public bid.

Each block producer $i \in \mathcal{P}$ has its private copy of the set of outstanding transactions, known as *mempool*. Recall that in RTR-TFM the block producers store both the transaction bids and the specific hop at which they first hear the transaction. That is, producer i's mempool is the tuple $\mathcal{M}_i = (F_i, H_i)$. $\mathcal{M}_i$ comprises the set of user bids $F_i = (b_1, \ldots, b_n)$ and their corresponding hops $H_i = (h_{i,1}, \ldots, h_{i,n})$.

TFM: Allocation (x), Payment (p), and Burning Rule (δ) [? ?]. To create its block, a block producer $i \in \mathcal{P}$ selects a subset of transactions from $S \in F_i$ to add. Each such *allocated* transaction $t \in S$, is charged a *payment* p_t for its slot in the block. E.g., if the TFM is a first-price auction, $p_t = b_t$. If the TFM is a second-price auction, then p_t equals the highest losing bid. Furthermore, TFMs can also include a *burning* rule, where a fraction of the payment p_t, say δ_t, is removed from the cryptocurrency's supply forever.

Strategy Space Trivially, the strategy space for users in a TFM is their bid, i.e., for each user $j \in \mathcal{U}$, the strategy space is (b_j). Typically, the block producer's strategy space comprises the allocation rule it picks, say x, and the choice of whether to add any fake transactions to blocks to manipulate the fee levels in the block, say $\mathcal{F}$. As payments are issued for routing transactions, we enrich the strategy space for producers to include the choice of whether to hoard transactions or *sybil* the routing path by adding false identities controlled by the attacker.

Utility Model [? ?].

The utility model for both users and block producers reflects the value that each class of participant can extract from the mechanism minus the costs paid for securing the related benefits. For each user $j \in \mathcal{U}$ with valuation θ_j and bid b_j we have the following utility structure:

$$u_j^{\mathcal{U}}(\theta_j, \{\theta_{j'}\}_{j' \in \mathcal{U} \setminus \{j\}}; p_j) := \begin{cases} (\theta_j - p_j - \delta_t) & \text{if} j \in S \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For each block producer we introduce $\mathcal{F}_i$, which is the set of fake transactions the block producer $i \in \mathcal{P}$ may include to strategically deviate from honest behavior. Now the producer's utility structure is:

$$u_i^{\mathcal{P}}(F_i, \mathcal{F}_i) := \underbrace{\sum_{t \in F_i} p_t}_{\text{Revenue from Honest Txs}} - \underbrace{\sum_{t' \in \mathcal{F}_i} \delta_{t'}}_{\text{Cost of Fake Txs}} \tag{2}$$

As previously mentioned, in a TFM mechanism designers consider incentive compatibility for both users and producers. It follows that:

Definition 5.1 (Incentive Compatibility Users [? ?]). We say that a TFM is incentive compatible for users if bidding their true valuation maximizes the utility of users, irrespective of the other user bids. That is, for any user $i \in \mathcal{U}$, $\forall \theta_j$ and $\forall \{\theta_{j'}\}_{j' \mathcal{U} \setminus \{j\}}$, we have

$$u_j^{\mathcal{U}}(b_i^{\star} = \theta_j, \cdot; \cdot) \geq u_j^{\mathcal{U}}(b_i = \theta_j, \cdot; \cdot).$$

Definition 5.2 (Incentive Compatibility - Producers). We say that a TFM is incentive compatible for block producers if it is the block producer $i \in \mathcal{P}$'s best response to (i) follow the intended TFM allocation rule and (ii) not include any fake transactions, i.e., $\mathcal{F} = \emptyset$, given that the remaining producers $\mathcal{P} \setminus \{i\}$ follow the same strategy.

6 CONCLUSION & FUTURE WORK

CONCLUSIONS HERE

In this paper, we introduced RTR-TFM: a novel TFM that addresses the incentive misalignment in classic transaction fee mechanisms (TFMs) by introducing a novel routing-based block production rule and a revenue scheme. RTR-TFM rewards block producers in proportion to their contribution to the propagation of transactions. Such a reward ensures that block producers actively participate in the blockchain network upkeep instead of free-riding on other participating nodes. We also provide a game-theoretic characterization of the underlying game in RTR-TFM. We prove that RTR-TFM effectively discourages transaction hoarding, ensures Sybil resistance, and achieves incentive compatibility for both users and block producers under reasonable assumptions.

Future Work. With RTR-TFM, we introduce a TFM revenue rule that links a direct cost to the block producers to create fake transactions. However, our BPIC analysis assumes a bootstrapped blockchain (Assumption **??**). While Assumption **??** is practical, we can look towards BPIC guarantees without constraints on the blockchain state. Furthermore, the TFM literature also looks at off-chain collusion between users and producers. [?] show the impossibility of a deterministic TFM simultaneously satisfying the UIC, the BPIC, and the resistance to off-chain collusion between a user and a producer. Future work can also study off-chain collusion guarantees for RTR-TFM.

Temporary page!

LaTeX was unable to guess the total number of pages correctly. As there was some unprocessed data that should have been added to the final page this extra page has been added to receive it.

If you rerun the document (without altering it) this surplus page will go away, because LaTeX now knows how many pages to expect for this document.