Achieving Incentive Compatibility with Collusion-Proof Equilibria

David Lancashire
david.lancashire@gmail.com

February 27, 2025

Abstract

A number of recent papers in computer science have applied mechanism design techniques to the study of collusion within transaction fee mechanisms (TFMs), and claim the results show the impossibility of building consensus mechanisms that are incentive compatible with collusion-free and other desirable equilibria. This paper identifies a specific methodological flaw in these results through the application of a composite utility model that endogenizes the costs and benefits of collusion and shows under what specific conditions it is rational for participants. This endogenous model is then used to identify the specific social choice rule needed to achieve incentive-compatibility with collusion-free outcomes and shows why previous models are incapable of preventing collusion: they fail to handle truthful preference revelation in the way required by implementation theory. These findings have significant implications for blockchain fee design, and suggest that appropriately structured TFMs can overcome the impossibility results dominating the field.

1 Introduction

Transaction Fee Mechanisms (TFMs) are a type of distributed system in which a consensus mechanism governs the allocation of the same resource that incentivizes its own provision. Unlike traditional consensus mechanisms, where the number of honest and dishonest participants is static, TFMs feature dynamic voting power that flows with the payouts issued by the mechanism.

The potential for users and producers to collude in ways that manipulate payouts has encouraged researchers to apply economic techniques to analyze whether collusion-free TFMs can be designed. In mechanism design terms, this means examining whether a mechanism can be incentive-compatible with a collusion-resistant equilibrium. Unfortunately, this research has produced a series of impossibility results suggesting that collusion-proof TFMs are infeasible.

This paper revisits these claims and shows they are based on methodological inconsistencies that lead directly to negative findings. Among the issues are: (1) the failure to model collusion as an endogenous problem with costs and benefits internal to the mechanism, (2) problems with inadequate and untruthful revelation of agent preferences, and (3) an implicit reliance on a social choice rule that fails to demand the high-dimensional preferences needed to compute collusion-free equilibria.

In order to demonstrate why these problems hold, our next section reviews how previous papers have modelled collusion and why these models claim incentive-compatibility cannot be achieved. Following that, we develop a composite utility model that makes the costs and benefits of collusion endogenous to the mechanism, and use it to identify which private preferences must be revealed to any incentive compatible mechanism. We then apply our composite utility function to the question of whether collusion-free equilibria are theoretically computable, and show that solutions exist with indirect mechanisms that target *pareto optimality* as their social choice rule.

These findings suggest that incentive-compatible and collusion-resistant equilibria are not only possible to implement, but will emerge naturally once mechanisms are structured to endogenously account for the costs and benefits of collusion.

1.1. Methodological Assumptions in the Existing Literature

Attempts to study blockchains using auction design techniques started with Lavi et al. (2017), which proposed Bitcoin shift to a "monopolistic mechanism" in which all transactions pay the same price for fee-stability in the face of a falling block reward. This was followed by an analysis from Yao (2018) which used implementation theory to argue the approach was "nearly" incentive compatible under atomistic conditions.

Early papers could assume users were purchasing a simple private good with an uncomplicated utility function – they focused on stabilizing miner revenue. But this led the assumption to carry over into later papers, where the choice to pay fees on-chain or off-chain suggested more complicated preferences were involved:

$$u_t\left(b_t\right) := \begin{cases} \left(v_t - p_t\left(H, B_k\right) - q_t\right) \text{ if } j \in S, \\ 0 \qquad\qquad\qquad\qquad\qquad\quad \text{otherwise.} \end{cases}$$

In this equation from Roughgarden (2024), one of the first papers to assert that collusion-proof TFMs are impossible to build, the utility to the user is modelled as the private value of blockspace, minus any payment from the user to the producer, minus any portion of the fee that is burned. This approach, which treats collusion as exogenous to the model, simplifies the task of computing viable equilibria, and has been adopted

by almost every paper written about collusion-resistance within TFMs (see Chen et al. (2022); Ferreira et al. (2021); Wu et al. (2023); Damle et al. (2024); Gafni and Yaish (2024); Bahrani et al. (2023, 2024); Chung et al. (2024); Chung and Shi (2023)).

A consequence is the belief incentive compatibility simply requires "truthful preference revelation" from users about the value of blockspace and "faithful implementation" of the mechanism by block producers. Truthful preference revelation is assumed to involve users bidding the highest fee they are theoretically willing to pay for blockspace in any strategic environment Roughgarden (2024); Chung and Shi (2023), a decision that asserts a user's maximum bid -— not their actual bidding strategy -— is the relevant signal of their truthful private preference.

This assertion sets the TFM literature at odds with the broader literature on implementation theory in economics, where truthful preference revelation is more rigorously defined. In their prize-winning work, Hurwicz and Maskin emphasize that truthful preference revelation requires agents to disclose all private information that affects strategy formation. Hurwicz refers to these as the "preference maps" of his agents, while Maskin describes them as the "characteristics" or "types" that must be revealed for a mechanism to successfully implement a social choice rule Hurwicz (1973, 1960, 2007, 1979); Maskin (1999, 2002).

When faced with the question of why only the highest-possible-bid represent truthful preference revelation, Ethereum-focused researchers like Roughgarden and Shi simply point to mechanisms in which the assumption is reasonable, such as the second-price Vickrey auction run by a trusted auctioneer. By comparing on-chain and off-chain implementations of similar auction mechanisms, these papers assume that if a bid constitutes adequate preference revelation in a mechanism with a trusted auctioneer then it should also be truthful in the presence of a potentially adversarial counterparty. The user is still reporting their utility of getting the good, no?

This argument is flawed because it assumes the truthfulness is a property of a bid that exists independently of the rationality of collusion – collusion becomes exogenous to the model. Most papers seem unaware this constitutes a problem, but it is incorrect to assume that a transaction fee that is truthful in a one-party game must be truthful in a two-sided auction. In Section 1.3 we will go into more detail on this subject and show how an endogenous model reveals the need for more high-dimensional preference revelation.

In any event, having asserted that the highest potential bid constitute truthful preference revelation, the aforementioned papers claim the impossibility of achieving incentive compatibility by finding situations in which threshold users still have incentives to under-bid, such as through bid-shading strategies. Alternately, a lack of incentive compatibility is demonstrated by pointing out that producers can manipulate fee-levels by creating fake transactions which replace any price-setting bid even if users do bid truthfully.

A minority of papers attempt more mathematically sophisticated attempts to achieve incentive compatibility by employing Bayesian models and techniques such as Myerson's Lemma to argue that stable equilibria may exist if users base their bidding strategies on historical data showing the density and distribution of transaction fees that were successfully included in the blockchain. Chen et al. (2022) offers one such paper that argues a resulting equilibrium is findable and should be considered collusion-proof. Unfortunately, the same problem affecting the non-Bayesian papers also affects the conclusions of this paper.

As the remainder of our own paper will demonstrate, mechanism design does not permit the assumption that the transaction fee constitutes a truthful revelation of preferences in direct mechanisms where there is a potential for user-producer collusion. As soon as our social choice rule requires agents to prefer collusion-free outcomes, a composite utility function is required that demands higher-dimensional preference revelation from users and producers than is possible to encode in a single transaction fee. And producers must have the strategic flexibility to signal their own relevant preferences.

To clarify these points, the following section presents a composite utility model that explicitly incorporates the benefits and costs of collusion. By making these factors endogenous to our analysis, we identify the precise conditions under which users will collude with producers. This allows us to identify the exact set of private preferences that incline users and producers to prefer not colluding to colluding. This approach will make it clear the transaction fee cannot constitute truthful preference revelation in the way expected by the existing TFM literature.

1.2. A Composite Utility Model for Blockchain

Accounting for user-producer collusion requires a composite utility model that accounts for both on-chain and off-chain payments. We define **public fee** as the portion of a transaction fee that is tendered openly for the competitive inclusion of the transaction in the blockchain and **private fee** as any portion distributed privately as an off-chain payment for the same good. The price p paid by user j is the sum of their public and private fees.

$$p_j = p^j_{pub} + p^j_{priv}$$

Users can purchase transaction inclusion using either a *public fee* or a *private fee*. *Private fees* are more appealing to producers because a larger portion of the fee can be extracted as profit in conditions of non-atomistic competition, where the shift to a private off-chain payment reduces the potential income available to other producers and diminishes their willingness to drive up the cost of any block production function like hashing or staking.

Whenever producers can extract a greater percentage of the overall fee as profit, they shift the allocation of the fees collected by the network into the production of alternate forms of utility. To model the impact of this shift on users, our composite utility function will eventually include three specific types of utility-providing goods: **public goods**, **private goods**, and **collusion goods**.

Our first category is *public goods*, which consist of non-excludable benefits that scale monotonically with the *public fees* included in a block. In non-atomistic conditions where producers are not compelled to maximize their spending on the security function, one such good is the economic security of the network. Other benefits commonly attributed to "decentralization" in TFMs are also *public goods* that scale with the volume of *public fees*, such as the degree of censorship resistance of the network and the ease with which new nodes can join it.

Our second category is *private goods*, which consist of the private benefits of transaction inclusion in the blockchain. The defining feature of this category is not its manner of funding, since users who purchase blockspace with *public fees* also accrue these benefits. Notably, this category does not include any additional benefits that accrue to users as a result of colluding with producers, since those benefits are not provided to non-colluding users who purchase blockspace with *public fees*.

In the absence of collusion, our valuation θ_j is the sum of the utility offered by these public and private goods:

$$\theta_j = U^j_{pub} + U^j_{priv}$$

Since the utility of the public good component scales with the total amount of public fees in the block, and the utility provided by private goods scales with the fees contributed exclusively by the fee-paying user, our valuation function becomes:

$$\theta_j = f^j_{pub}\left(\sum_{k\in S} p^k_{pub}\right) + f^j_{priv}(p^j_{priv})$$

Which gives us our full utility function in the absence of collusion:

$$u^U_j\,(...) := \begin{cases} \left(f^j_{pub}\left(\sum_{k\in S} p^k_{pub}\right) + f^j_{priv}(p^j_{priv})\right) - \left(p^j_{pub} + p^j_{priv}\right) & \text{if } j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

To incorporate the costs and benefits of user-producer collusion in our model, we add these elements to our equation. *Collusion goods* refer to any benefits producers offer users in exchange for a *collusion fee* whose name indicates it will be allocated by the producer to the production of this type of utility.

Our updated valuation function becomes:

$$\theta_j = f^j_{pub}\left(\sum_{k\in S} p^k_{pub}\right) + f^j_{priv}(p^j_{priv}) + f^j_{col}(p^j_{col})$$

This equation illustrates the dynamics of collusion in TFMs. Users control how fees are offered to producers, and their choice affects the amount of profit producers can extract and the degree of competition they face to produce blocks. This in turn determines the degree of flexibility producers have in allocating fees to the production of different kinds of utility, which affects the utility producers can offer to users in exchange for private control of their transactions.

For users, the valuation function reflects the sum of the utility provided by *public goods*, *private goods* and *collusion goods*. The attractiveness of collusion is determined not only by their private valuation for transaction inclusion (as is assumed in the univariate models discussed in Section 1.1) but also the potential availability and cost of *collusion goods* and the degree to which other users allocate their fees in ways that lead to the funding of *public goods*.

Since the utility provided by *private goods* is beyond the ability of producers to manipulate – given to all transactions in the blockchain regardless of their form of payment – the rationality of collusion for users depends on the comparative marginal utility offered by the other two forms of utility in play: *public goods* and *collusion goods*.

This lets us define collusion as any cooperative action in which users and producers re-allocate any portion of a *public fee* to a *collusion fee*. This definition accommodates exceptions that arise within the model. Any increase in the provision of *collusion goods* that does not decrease the level of *public fees* is strictly utility-increasing and can be viewed as a voluntary trade exogenous to the model. Any reduction in the *public fee* that leads to an increase in total fees paid can be modeled as an act of collusion coupled with a separate voluntary trade.

In situations where collusion results in a discounted overall fee or cash refund, we treat the value of the cash discount as the utility offered by the *collusion good*.

Our price shift under collusion becomes:

$$p_j = \left(p^j_{pub} - p^j_{col}\right) + \left(p^j_{priv}\right) + \left(p^j_{col}\right)$$

And our utility shift becomes:

$$\theta_j = f^j_{pub}\left(\sum_{k\in S} p^k_{pub} - p^j_{col}\right) + f^j_{priv}\left(p^j_{priv}\right) + f^j_{col}\left(p^j_{col}\right)$$

Since our total fee is unchanged, collusion is attractive if the fee-shift increases utility, i.e.:

$$f^j_{pub}\left(\sum_{k\in S} p^k_{pub} - p^j_{fr}\right) + f^j_{priv}(p^j_{priv}) + f^j_{col}\left(p^j_{col} + p^j_{fr}\right) > f^j_{pub}\left(\sum_{k\in S} p^k_{pub}\right) + f^j_{priv}(p^j_{priv}) + f^j_{col}\left(p^j_{col}\right)$$

This equation makes it clear the transaction fee cannot – on its own – constitute adequate preference revelation.

For collusion to be rational the marginal utility of at least one *collusion good* must be higher than the marginal utility of the *public goods* competing for consumption of the same fee. Since our *collusion good* has the implied utility of cash, it follows that collusion is rational in any situation where the marginal utility of *any other good* is higher than the marginal utility of the *public goods* provided by the TFM.

The relevant private preferences that must be revealed to any direct mechanism thus constitute the comparative marginal utility of multiple goods to fee-paying users, and their cost-of-production to producers. Collusion becomes rational when users value *collusion goods* more than *public goods* and/or producers have cost advantages in producing *collusion goods* which make them more efficient providers of utility than can be purchased with a *public fee*.

1.3 Methodological Problems in the Computer Science Literature

As discussed in Section 1.1, the vast majority of attempts to model collusion within TFMs treat blockspace as if it is a private good with a non-composite form of utility, making collusion exogenous to the model. This approach simplifies the work needed to calculate viable equilibria by reducing the problem to what Hurwicz called a "one-objective maximization function" but obscures the motivations driving users to collude. As such, it misleads us regarding the preferences that must be revealed to any mechanism seeking incentive compatibility with a collusion-free outcome.

We can now see this approach creates methodological contradictions.

The first issue arises from the assumption that a bid in a TFM constitutes "truthful preference revelation" simply because a mechanism allocates blockspace using a pricing algorithm in which bids are deemed truthful in another context. As per Maskin, any change in social choice rule necessitates a reconsideration of what constitutes a relevant preference. In a Vickrey auction, the auction mechanism's social choice rule is the "efficient allocation" of a single good, meaning that the transaction fee need only encode users' comparative preference for that single form of utility. Here, our switch to requiring users and producers to prefer non-collusion outcomes expand the scope of relevant preferences to include the comparative utility users and producers can gain from re-allocating a portion of their *public fees* to the purchase of *collusion goods*.

The assertion that incentive compatibility requires producers to "faithfully implement" mechanisms is also untenable, for our endogenous model shows that achieving incentive compatibility in fact requires the opposite – that producers must directly or indirectly reveal their cost basis to the mechanism. The axiomatic refusal to consider mechanisms in which producers reveal this information – such as through the selective and strategic inclusion of their own transaction fees in blocks – becomes a prima facia reason exogenous models are incapable of targeting collusion-proof outcomes.

Fascinatingly, the Bayesian approaches mentioned in Section 1.1 also collapse in the face of this problem with inadequate preference revelation, since it stops being possible to invoke the Revelation Principle and Myerson's Lemma to generalize about "collusion-free" bidding strategies if the historical bids whose distribution and/or density are used to calculate viable equilibria reflect bidding strategies in which collusion is rational and bids are consequently untruthful.

Importantly, none of these points are intended to suggest that the transaction fee is theoretically incapable of constituting adequate and truthful preference revelation within the theoretical framework advanced by Hurwicz and Maskin. It remains possible that a transaction fee may still constitute truthful preference revelation in TFMs should it encode the relevant preferences obliquely, such as might occur if users only submit bids once they have privately calculated that collusion is suboptimal. In this case the existence of the bid would suffice to indicate the marginal utility of the *public good* is higher to the user than the marginal utility of any *collusion good*. In the language of mechanism design, we would have shifted from a "direct mechanism" to an "indirect mechanism".

Unfortunately, even if we consider indirect mechanisms the conclusions of the aforementioned papers cannot hold since they leverage the Revelation Principle and Myerson's Lemma to generalize their findings. Yet Myerson's Lemma is dependent on the Revelation Principle, and the Revelation Principle only permits generalization from indirect mechanisms to direct mechanisms. Impossibility results cannot be generalized

in the opposite direction, since the underlying logic of the Revelation Principle is the observation that a symmetry of outcomes must exist between mechanisms where information is computed in decomposable fashion using agent-level functions, and centralized mechanisms where the exact same information is revealed truthfully and the computation is performed identically in a non-decomposable fashion. The fact that indirect mechanisms work using a broader slate of preferences than are observable to direct mechanisms make it impossible to generalize impossibility proofs in the other direction, since we cannot conclude that solutions do not exist which exploit information and techniques unobservable to our models.

In short, without truthful preference revelation we cannot use the standard tools of mechanism design used in the TFM literature to generalize about the possibility or impossibility of achieving incentive compatibility in two-sided auctions in which participants have potentially rational opportunities to collude. While this finding is negative in the sense that it offers a general critique of most of the existing literature, the shift to composite utility functions also opens the door to solutions that are otherwise obscure, and points directly to a specific subclass of mechanisms in which a solution seems discoverable: indirect mechanisms which implement pareto optimality as their social choice rule.

In order to show why this is the case, in the next section we apply the composite utility function to the analysis of this social choice rule – the only state of equilibrium we can guarantee to be collusion-free – and find a surprising affinity with another known and known-solvable problem in the field of welfare economics.

1.4. Pareto Optimality, Free-Riding and Collusion-Free Equilibria

In order for collusion to be rational, users and producers must be able to profitably re-allocate resources to collusion goods. The one condition in which this is provably irrational is when production is already on the *utility possibilities frontier* and all participants are already allocating their resources in whatever way maximizes their private utility. From this state of production, all forms of collusion are necessarily irrational as it is not possible for any subset of participants to adjust the way their own resources are allocated without making at least one member worse off. Were such a shift possible, we would by definition not be on the *utility possibilities frontier*.

Achieving production on the *utility possibilities frontier* is characteristic of a *pareto optimal* equilibrium, which is why *pareto optimality* must be the social choice rule implemented by any mechanism seeking incentive compatibility with a collusion-free equilibrium.

To express this social choice rule mathematically, we introduce two cost functions $F_{pub+priv}()$ and F_{col} to express the costs of producing our competing forms of on-chain (public and private) and off-chain (collusion) utility. As per the following equation, *pareto optimality* is achieved when the marginal utility per unit of cost is equalized across all possible allocations, ensuring that fees cannot be reallocated across *public goods*, *private goods* and *collusion goods* to increase the total amount of utility produced.

$$\sum_{j=1}^{s} \frac{u^{j}_{pub+priv}}{u^{j}_{col}} = \frac{F_{pub+priv}}{F_{col}}$$

Economists will recognize this as structurally identical to the equation Samuelson (1954) flagged in his seminal paper on the difficulty of inducing free markets to produce non-excludable goods in *pareto optimal* amounts. What the symmetry shows is that free-riding pressures create conditions for collusion: the non-excludable nature of the *public goods* in the blockchain encourages rational actors to underfund them. And because this pulls the network out of any *pareto optimal* equilibrium, it pulls production off the *utility possibilities frontier*. Collusion becomes rational because the ability of users to free-ride pulls us out of the only known collusion-free state.

The challenges of designing collusion-proof TFMs thus go beyond the implementation problems identified by Hurwicz in 1972; they also require solving the deeper inefficiencies outlined by Samuelson two decades earlier. Without truthful preference revelation, optimal outcomes can be subverted by strategic misdirection from agents who circulate false preference maps. But even if we have a mechanism with an "incentive to truthfulness", free-riding pressures will still destabilize optimal equilibria unless they are explicitly neutralized. Both problems must be eliminated to build TFMs that are incentive compatible with collusion-resistant equilibrium.

Understanding the link between the rationality of free-riding and the rationality of collusion does more than highlight the analytic limitations of non-composite utility models which obscure the connection – it also suggests practical solutions. For an example of this, observe that auction mechanisms which give producers a "temporary monopoly" over blocks are necessarily vulnerable to free-riding as the probability of producing any block becomes disconnected from the volume of fees included in the block. Auction mechanisms that grant producers monopolistic "slots" to propose blocks necessarily empower them to extract fees without reinvesting in network security.

While this specific problem may seem intractable to those accustomed to *proof-of-stake* mechanisms, there are no compelling reasons to considering the issue theoretically unsolvable: solutions already exist. One method of eliminating it is to make the speed of block production dependent on the volume of fees included in the block, as in routing work mechanisms that pair a Dutch clock auction with a descending fee-burn. In situations where the threshold user attempts to engage in bid-shading, the amount of work in the block is reduced and – with it – the competitiveness of the block producer in purchasing the next block. Producers who manipulate fee-volumes are likewise forced to add their own tokens to a fee-burn, which provides a basis for asymmetrically punishing their own attempts at fee-manipulation.

A second and more unexpected technique that eliminates free-riding is explicitly incentivizing producers to share unconfirmed transactions with their peers. This second strategy exploits the same principle as the atomistic market exception: in any situation where producers are forced to compete intensely for the right to produce blocks, producers cannot reduce the percentage of their income allocated to block production. And without the ability to reallocate fees to the production of any *collusion good*, colluding with users becomes irrational. By incentivizing the sharing of unconfirmed transactions, mechanisms can approximate the competitive intensity of atomistic markets, in which rational producers will not collude with users even if users desire it.

We can model this second approach mathematically by simplifying our cost function to require only the payment of a single fee and adding a variable x that reflects the probability that transactions are circulating publicly and available for competitive inclusion. Normalizing x into a number between 0 (non-atomistic competition) and 1 (atomistic competition) gives us:

$$\sum_{j=1}^{s} \frac{u^j_{(pub*x)+priv}}{u^j_{col}} = \frac{F_{priv}}{F_{col}}$$

As the value of x approaches 1, free-riding becomes irrational. To understand why, observe the utility provided by the TFM itself:

$$u^j_{(pub*x)+priv} = F_{priv}$$

Users and producers have different preferences for the value of x:

- **rational users** - prefer x to approach 1, to maximize competition for transaction inclusion.
- **rational nodes** - prefer to free-ride on publicly-circulating transactions but not share their own.

Since producers will not discriminate against transactions bearing *public fees* ceteris paribus, the obstacle to achieving maximally-intensive intra-producer competition is the strategic preference of block producers to limit competition for collection of the fees in their mempool. Reverse those incentives and producers will prefer to share transactions with their peers, permitting x to approach 1 and ensuring the marginal utility provided by the blockchain increases monotonically with total fees paid.

This shows a mathematical connection between the problems of collusion within TFMs and the *sybil problem* identified by Babaioff et al. (2011). A theoretical claim it is impossible to solve this problem in *proof-of-work* and *proof-of-stake* networks may be found in the above-cited paper *On Bitcoin and Red Balloons*. We nonetheless observe a subset of routing work mechanisms that can address this specific problem. As such, at least a subset of indirect mechanisms exists which avoid the Samuelson suboptimality trap by simplifying his equation to the following once x becomes 1:

$$\sum_{j=1}^{s} \frac{u^j_{pub+priv}}{u^j_{col}} = \frac{F_{priv}}{F_{col}}$$

While this approach does not ensure that fee-levels will achieve *pareto optimal* levels, it ensures that if a mechanism can otherwise implement *pareto optimality* as its social choice rule, incentives for free-riding will not drag the mechanism out of its collusion-free equilibrium and back into a situation in which collusion may be rational for a subset of network participants.

While hardly comprehensive, these approaches suggest at least two viable approaches for making progress towards collusion-resistant equilibrium: eliminating the ability for mechanisms to offer "temporary monopoly" to block producers, and incentivizing producers to share unconfirmed and fee-bearing transactions with their peers.

1.5 Conclusions

The shift to modeling collusion using a composite utility function allows for a better understanding of the incentives driving collusion in transaction fee mechanisms (TFMs).

First, it reveals significant methodological flaws in prior work on TFMs. By making the incentives for collusion endogenous to the utility model, we demonstrate that transaction fees alone can never constitute adequate and truthful preference revelation in any direct mechanism aiming for incentive compatibility with a collusion-free outcome. This directly undermines the impossibility claims made in many earlier papers. Given the tendency for academic research to build upon prior results, this finding underscores the need for a return to more rigorous foundational work.

Second, our composite utility model shows that *pareto optimality* is the only social choice rule capable of eliminating collusion. By definition, it is only on the *utility possibilities frontier* that the marginal costs of all forms of competing utility are brought into alignment. Whether our goal is to prevent users from bid-shading or discourage producers from using their own money to produce blocks, the *pareto optimal* equilibrium is the only viable outcome for accomplishing it.

Third, composite utility models highlight the presence of free-riding pressures in TFMs, and show how the existence of these problems within a TFM can pull the network out of the only equilibrium in which collusion is irrational. This shows that efforts to achieve incentive compatibility must involve more than achieving the "incentive to truthfulness" identified by Hurwicz in 1972. Without eliminating the kinds of free-rider

pressures identified by Samuelson as well, incentives will always exist for at least a subset of participants to collude in ways that benefit them but hurt overall social utility.

Beyond these theoretical insights, this paper offers three practical recommendations for academics working on the analysis of incentive compatibility in TFMs.

For computer scientists evaluating the incentive properties of consensus mechanisms, we show it is critical to explicitly define the social choice rule mechanisms seek to implement before drawing conclusions about their incentive compatibility. Identifying the targeted outcome allows for a rigorous assessment of which private preferences influence strategy formation, whether they are being revealed to the mechanism, and which tools from mechanism design are consequently appropriate for subsequent analysis.

For economists familiar with mechanism design, this paper highlights that the problems subverting socially optimal outcomes in TFMs expand beyond the problems with strategic manipulation studied by Hurwicz and Maskin. They also include the collective action problems identified by Samuelson and other public choice theorists. The fact that both problems appear solvable with blockchain-based mechanisms suggests that blockchains may be able to achieve socially-optimal outcomes in situations where free markets fail.

For game theorists, the presence of free-riding problems in TFMs suggests that more attention is needed to indirect mechanisms, since this class of mechanisms is traditionally the appropriate type for handling the higher-dimensional preference revelation needed to optimize the provision of public goods. As with the Clarke-Groves mechanism which uses bundled-bidding to induce multi-variate preference revelation, solutions capable of eliminating collusion are most likely to be found in mechanisms with decomposable algorithms where preference filtering is handled by users and producers prior to fee-selection.

On a final note, we close by observing that this paper provides not only predictive power but falsifiable claims, since it predicts that any technical shift that makes headway on the informational problems that subvert pareto optimal equilibria in informationally decentralized mechanisms should reduce the scope for collusion within TFMs. It also suggests that any viable solution to the Red Balloons sybil problem will transitively reduce any incentives that block producers have to collude with users.

References

Moshe Babaioff, Shahar Dobzinski, Sigal Oren, and Aviv Zohar. 2011. On Bitcoin and Red Balloons. *CoRR* abs/1111.2626 (2011). arXiv:1111.2626 http://arxiv.org/abs/1111.2626

Maryam Bahrani, Pranav Garimidi, and Tim Roughgarden. 2023. *Transaction Fee Mechanism Design with Active Block Producers*. Technical Report. arXiv. org.

Maryam Bahrani, Pranav Garimidi, and Tim Roughgarden. 2024. Transaction fee mechanism design in a post-mev world. *Cryptology ePrint Archive* (2024).

Xi Chen, David Simchi-Levi, Zishuo Zhao, and Yuan Zhou. 2022. Bayesian mechanism design for blockchain transaction fee allocation. *arXiv preprint arXiv:2209.13099* (2022).

Hao Chung, Tim Roughgarden, and Elaine Shi. 2024. *Collusion-Resilience in Transaction Fee Mechanism Design*. Technical Report. arXiv. org.

Hao Chung and Elaine Shi. 2023. Foundations of transaction fee mechanism design. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 3856–3899.

Sankarshan Damle, Manisha Padala, and Sujit Gujar. 2024. Designing Redistribution Mechanisms for Reducing Transaction Fees in Blockchains. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 416–424.

Matheus VX Ferreira, Daniel J Moroz, David C Parkes, and Mitchell Stern. 2021. Dynamic posted-price mechanisms for the blockchain transaction-fee market. In *Proceedings of the 3rd ACM Conference on Advances in Financial Technologies (AFT)*. 86–99.

Yotam Gafni and Aviv Yaish. 2024. Barriers to collusion-resistant transaction fee mechanisms. *arXiv preprint arXiv:2402.08564* (2024).

Leonid Hurwicz. 1960. Optimality and Informational Efficiency in Resource Allocation Processes. *The American Economic Review* 50, 3 (1960), 462–477.

Leonid Hurwicz. 1973. The Design of Mechanisms for Resource Allocation. *The American Economic Review* 63, 2 (1973), 1–30. http://www.jstor.org/stable/1817047

Leonid Hurwicz. 1979. On Allocations Attainable Through Nash Equilibria. *The Journal of Economic Theory* 21, 1 (1979), 140–165.

Leonid Hurwicz. 2007. But Who Will Guard the Guardians? Nobel Prize Lecture. https://www.nobelprize.org/prize-events/2007/ University of Minnesota, Department of Economics.

Ron Lavi, Or Sattath, and Aviv Zohar. 2017. Redesigning Bitcoin's Fee Market. *arXiv preprint arXiv:1709.08881* (2017). https://arxiv.org/abs/1709.08881

Eric Maskin. 1999. Nash Equilibrium and Welfare Optimality. *The Review of Economic Studies* 66, 1 (1999), 23–38. https://doi.org/10.1111/1467-937X.00086

Eric Maskin. 2002. Implementation Theory. In *Handbook of Social Choice and Welfare*, K. Arrow, A. Sen, and K. Suzumura (Eds.). North-Holland, 511–587.

Tim Roughgarden. 2024. Transaction Fee Mechanism Design. *J. ACM* 71, 4 (2024), 1–25.

Paul A. Samuelson. 1954. The Pure Theory of Public Expenditure. *The Review of Economics and Statistics* 36, 4 (1954), 387–389. https://doi.org/10.2307/1925895

Ke Wu, Elaine Shi, and Hao Chung. 2023. Maximizing miner revenue in transaction fee mechanism design. *arXiv preprint arXiv:2302.12895* (2023).

Andrew Chi-Chih Yao. 2018. An Incentive Analysis of some Bitcoin Fee Designs. *arXiv preprint arXiv:1811.02351* (2018). https://arxiv.org/abs/1811.02351 Version 3, last revised 11 Nov 2018.