Collusion and Free-Riding in Transaction Fee Mechanisms

David Lancashire
david.lancashire@gmail.com

February 2, 2025

Abstract

Recent papers in computer science have applied mechanism design techniques to the study of user-producer collusion within transaction fee mechanisms (TFMs). While many of these works claim to establish impossibility results, their conclusions result from logical inconsistencies in their application of mechanism design theory. This paper addresses these issues through the introduction a composite utility model that endogenizes the costs and benefits of collusion, and shows the flawed assumptions about truthful preference revelation. Building on this, we identify the specific social choice rule needed to achieve incentive-compatibility with collusion-free outcomes and discuss the challenges in implementing this rule. These findings have significant implications for blockchain mechanism design, suggesting that appropriately structured TFMs can overcome the impossibility results currently dominating the literature, but are likely to involve indirect mechanisms which are inadequately studied by existing papers.

1 Introduction

Transaction Fee Mechanisms (TFMs) are a class of distributed systems in which a consensus mechanism governs the allocation of the very 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 adjusts with the payouts issued by the mechanism. As a result, any form of collusion that lowers fee throughput or subverts the payout mechanism creates systemic security risks.

The potential for users and producers to manipulate fee levels through collusion 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 designing collusion-proof TFMs is infeasible.

This paper revisits these claims and demonstrates that the models used to support them contain methodological inconsistencies that lead directly to their negative findings. Among the issues are: (1) the assumption that truthful preference revelation is a defining condition of incentive compatibility rather than a property that emerges with aligned incentives; (2) the failure to model collusion as an endogenous trade-off with costs and benefits internal to mechanism; and (3) the reliance on a social choice rule that fails to capture the high-dimensional preferences relevant to agent strategy formation.

In order to demonstrate this, our next section starts with a review of how previous TFM papers have approached the challenge of modelling collusion, and shows why these approaches lead to pessimism on the feasibility of building incentive-compatible TFMs. Following that, we present a composite utility model that makes the costs and benefits of collusion endogenous to the mechanism, and examine this model to determine which private preferences must be revealed to any mechanism in the course of truthful preference revelation. After identifying the informational limitations this implies as regards previous research, we apply our composite utility function to ask if collusion-free equilibria are theoretically possible, and find that solutions are likely to exists as indirect mechanisms with decomposable algorithms that target *pareto optimality* as their social choice rule.

This suggests that incentive-compatible, collusion-resistant TFMs are not only possible but naturally emerge when mechanisms are structured to endogenously account for the costs and benefits of collusion.

1.1. Methodological Assumptions in the Existing Collusion Literature

Attempts to study Bitcoin using auction design techniques started with **?**, which proposed Bitcoin shift to a "monopolistic mechanism" in which all transactions pay the same fee to engender fee-stability in the face of the network's falling block reward. This was followed by an analysis from Yao (2018) which discussed the proposal using the concept of incentive compatibility and argued the approach was "nearly" incentive compatible under atomistic conditions.

An implicit assumption in these early papers was the appropriateness of modelling transaction inclusion should be modelled as private good. This assumption would become a central tenet of publications and encouraged researchers to define the utility blockchains offered users as their private valuation for blockspace minus any fees bid for securing a slot:

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

In this equation from Roughgarden (2024), in one of the first papers to explicitly discuss incentives for collusion in TFMs, the utility to the user is the value of transaction inclusion, minus any payment from the user to the producer, minus any portion of the fee that is burned. This modeling assumption, which treats collusion as exogenous to the model, has become characteristic of almost every paper written about collusion-resistance within the TFM literature (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)).

The TFM literature has also followed Roughgarden in defining incentive compatibility as requiring "truthful preference revelation" from users and "faithful implementation" of the mechanism by block producers. Truthful preference revelation assumed to involve users bidding the highest fee they are theoretically willing to pay for blockspace Roughgarden (2024); Chung and Shi (2023). This implicitly assumes that 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 curious 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 within a mechanism. 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).

Despite the foundational importance of these works, they are absent from most recent discussions on transaction fee mechanism design. The omission appears to stem from authors in the field starting from auction-specific definitions of truthfulness rather than conducting a first-principles examination of what private preferences are relevant to strategy formation in different mechanisms.

This omission is significant. When faced with the question of why higher bids represent truthful preference revelation, researchers like Roughgarden and Shi simply reference other mechanisms in which the assumption is considered reasonable, such as the second-price Vickrey-Clarke-Groves (VCG) auction run by a trusted auctioneer. By comparing an on-chain and off-chain implementation of otherwise identical pricing algorithms, these papers advance the unstated assumption that if a bid is deemed truthful and adequate preference revelation in a trusted auctioneer context then it should also be regarded as such in the presence of a potentially adversarial counterparty.

This assumption treats the truthfulness of a bid as a property that exists independently of the rationality of collusion – collusion becomes exogenous to the model. Many researchers do not seem aware they are pushing the incentives for collusion outside their model, yet this decision is methodologically critical. As the work of Hurwicz and Maskin illustrates, it is incorrect to assume that a bidding strategy deemed truthful in one context is automatically truthful in another. 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 the higher bid must constitute truthful preference revelation, the impossibility of achieving incentive compatibility is demonstrated in the TFM literature 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 shown by demonstrating that should users nonetheless bid truthfully, producers can manipulate fee-levels by creating fake transactions which replace the price-setting bid.

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 successfully purchased inclusion in the blockchain. Chen et al. (2022) offers one such paper that argues a resulting equilibrium is findable should be considered collusion-proof.

As the remainder of this 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 a preference for collusion-free equilibria, 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.

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

1.2. A Composite Utility Model for Collusion

Modelling collusion endogenously 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 paid by user j is the sum of their public and private fees.

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

While users can purchase transaction inclusion using either a *public fee* or a *private fee*, the utility they receive may differ based on both their and others' choices. *Private fees* are more appealing to producers because a larger portion of these fees can be extracted as profit in conditions of non-atomistic competition, where the privatization of any fee-competitive user transaction reduces the potential profits 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 fees collected by the network towards 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 associated with "decentralization" in TFMs also qualify as *public goods*, such as the degree of censorship resistance of the network and the lack of barriers of entry preventing new nodes from joining the network.

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, as those benefits are not issued 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 simply 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 intuitively illustrates the dynamics of collusion in TFMs. Users control how fees are offered to producers, and their choices affect the potential profits available to producers and the degree of competition they face to produce blocks. This interplay determines the degree of flexibility producers have in allocating fees to the production of different forms of utility, including those provided by *collusion goods*.

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 the inclusion of their own transaction (as is assumed in the univariate models) but also the potential availability of *collusion goods* and the way that other users and producers distribute and allocate their own fees to any *public goods* available from the chain.

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 they derive from *public goods* and *collusion goods*.

We 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 any exceptions that may arise within the model. Specifically, any increase in the provision of *collusion goods* that does not decrease the level of *public fees* is strictly utility-increasing for all participants and can be viewed as a voluntary trade exogenous to the model. Likewise, any reduction in the *public fee* that comes with 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 cost of transaction inclusion or cash refund, we simply treat the value of the discount as the utility offered by the *collusion good* being purchased.

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 re-allocation increases utility, i.e.:

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

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* can take the form of a discount or cash refund, 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* offered by the blockchain.

The relevant private preferences that affect the attractiveness of collusion thus compromise the comparative marginal utility of goods to our fee-paying users, and their cost-of-production to producers. Collusion is rational when users value *collusion goods* more than *public goods* and/or producers have cost advantages in the production of *collusion goods* which make them more efficient providers of utility than their peers.

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 utility function, 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 direct mechanism aiming for incentive compatibility with a collusion-free outcome.

We can now see this approach leads to fundamental 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 familiar pricing algorithm in which bids are deemed truthful in another context. Any change in social choice rule necessitates a reconsideration of what constitutes a relevant preference. In a Vickrey-Clarke-Groves (VCG) 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 non-collusive outcomes complicates preference revelation by expanding the scope of relevant preferences to include the comparative utility users can gain from re-allocating a portion of their *public fees* to the purchase of *collusion goods*.

Similarly, the assertion that incentive compatibility requires producers to "faithfully implement" a mechanism without acting strategically becomes untenable. Our endogenous model shows that achieving incentive compatibility in fact requires the opposite – that producers 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 – is a prima facia reason exogenous models are incapable of achieving collusion-proof outcomes.

Bayesian approaches also collapse in the face of this problem, 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 density are used to calculate bayesian nash equilibria reflect bidding strategies in which collusion is rational or bids are untruthful.

Importantly, none of these points are intended to suggest that the transaction fee is theoretically incapable of constituting adequate and truthful preference revelation. It remains possible that the transaction fee may still constitute truthful preference revelation if it is shown to encode the relevant preferences obliquely, such as might occur if users only submit bids in the event that they have already privately calculated that collusion is suboptimal. In this case the existence of the bid itself 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 this possibility the conclusions drawn by the aforementioned papers cannot hold since they attempt to leverage the Revelation Principle and Myerson's Lemma in order to advance their proofs. Myerson's Lemma is dependent on the Revelation Principle, and what the Revelation Principle

only permits generations from indirect mechanisms to direct mechanisms. Impossibility proofs cannot be generalized in the opposite direction, since Maskin's proof was developed 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. The fact that indirect mechanisms may work on a broader slate of preferences than those considered by direct mechanisms make it impossible to conclude that solutions to otherwise intractable problems do or not exist in that part of the solution space.

In short, without truthful preference revelation we cannot use the standard tools of mechanism design as used in the study of mechanisms with trusted auctioneers 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 may seem negative in that it offers a general critique of most of the existing literature, the shift to use of composite utility functions opens the door to progress on fundamental problems, by pointing directly to the subclass of mechanisms in which a solution seems likely to be found: 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 our composite utility function to the analysis of this social choice rule – the only one 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. A Collusion-Free Equilibrium in Composite Utility Models

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 on the *utility possibilities frontier* where all participants are already spending their resources in whatever way maximizes their own utility. This makes collusion 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*.

This condition characterizes 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 free markets producing 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 network 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 free-riding subverts efficiency.

The challenges of designing collusion-proof TFMs go beyond the implementation problems identified by Hurwicz in 1972; they also require solving the deeper inefficiencies outlined by Samuelson nearly two decades earlier. Without truthful preference revelation, mechanisms can be pulled out of optimality by the strategic manipulation of agents circulating false preference maps. But even if we create a mechanism with an "incentive to truthfulness", free-riding pressures will still destabilize it unless they are explicitly neutralized. Both problems must be eliminated to achieve incentive compatibility with a collusion-resistant equilibrium.

Understanding the need to eliminate free-riding does more than highlight the analytic limitations of non-composite utility models – 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 granting producers monopolistic "slots" to propose blocks thus empower them to extract fees without reinvesting in network security.

While this problem may seem intractable to those accustomed to the limitations of *proof-of-stake* mechanisms, there are no compelling reasons to considering the issue unsolvable. One method of eliminating it is to make the speed or cost 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 contribute their own tokens to a fee-burn, which provides a basis for asymmetrically punishing attempts at fee-manipulation.

A second approach to eliminating free-riding is explicitly incentivizing producers to share unconfirmed transactions with their peers. This strategy exploits the same principle as the atomistic market exception: in any situation where producers are forced to compete equally for the right to produce blocks and collect

payments, producers cannot reduce the percentage of their income allocated to block production. And without the ability to reallocate fees away from the *public good*, colluding with users to free-ride on *public fees* becomes irrational. By making the sharing of confirmed transactions profitable, we approximate the conditions of atomistic markets, where rational producers will not collude with users to privatize access to their transaction fees.

We can model this 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. We then normalize x to be a number between 0 (non-atomistic competition) and 1 (atomistic competition) and get:

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

As the value of x approaches 1, the utility provided by the blockchain scales monotonically with the total fees paid for transaction inclusion. To understand why this approach mitigates free-riding, observe the preferences of users and producers have to share their transactions publicly:

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

We observe split 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 eliminating free-riding in conditions of *pareto optimality* is the strategic preference of block producers to limit competition for collection of the fee. Reverse those incentives and producers will also prefer to share transactions with their peers, permitting x to approach 1 and forcing the public utility provided by the blockchain into alignment with the private utility of transaction inclusion.

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*. There is nonetheless evidence a subset of routing work mechanisms can indeed address this problem by incentivizing producers to share unconfirmed transaction flow with their peers. As such, at least a subset of indirect mechanisms exists which avoid the Samuelson suboptimality trap by simplifying his to the following once x becomes 1:

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

Once the marginal cost of utility from *private goods* and *collusion goods* is in alignment, free-riding that disturbs this equilibrium is no longer rational. 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 this equilibrium and back into a situation in which collusion may be rational for a subset of network participants.

Eliminating sybilling pressures by using routing payouts to induce the cooperative sharing of transactions is theoretically akin to having an atomistic market structure.

1.5 Conclusions

The shift to modeling collusion through a composite utility function is essential for understanding and addressing the incentives that drive collusion within transaction fee mechanisms (TFMs). This approach provides key insights that challenge existing literature and offer new directions for research.

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 truthful preference revelation in any direct mechanism aiming for incentive compatibility with a collusion-free outcome. This directly undermines the impossibility claims made in earlier papers, many of which rely on flawed assumptions. 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 utility for all goods relevant to the strategy formation are brought into an alignment where socially-suboptimal strategies are collectively irrational. 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 one needed to accomplish it.

Third, this approach shows that free-riding pressures in TFMs create conditions under which collusion becomes rational. Efforts to eliminate collusion soely through

ws that without addressing collective action problems, incentive compatibility remains unstable. Efforts to eliminate collusion solely through strategic mechanism design—such as those building on Hurwicz's

work—are insufficient unless they also account for free-riding incentives. Collusion arises whenever participants have the ability to shift the allocation of network resources in ways that privately benefit them at the expense of overall social utility.

Third, composite utility models highlight the presence of free-riding pressures in TFMs, and show how this class of collective action problems create the conditions under which collusion can become rational. This shows that efforts to achieve incentive compatibility by solving the set of problems identified by Hurwicz in 1972 are incapable of achieving a stable. Without eliminating free-riding pressures, incentives will always exist for at least a subset of participants to collude in ways that shift the overall allocation of network resources in ways that benefit them while reducing overall social utility.

Beyond these theoretical insights, this paper offers three key recommendations for advancing research on incentive compatibility in TFMs:

For computer scientists evaluating the incentive properties of consensus mechanisms, it is critical to explicitly define the social choice rule the mechanism seeks to implement before drawing conclusions about its 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 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. It predicts that any technical shift that makes headway on the fundamental informational problems that subvert achieving pareto optimal outcomes in informationally decentralized mechanisms should reduce the scope for collusion within TFMs. It also predicts that any viable solution to the Red Balloons sybil problem will transitively eliminate free-riding pressures.

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.

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.