On Collusion

Special thanks to Glen Weyl, Phil Daian and Jinglan Wang for review

Over the last few years there has been an increasing interest in using deliberately engineered economic incentives and mechanism design to align behavior of participants in various contexts. In the blockchain space, mechanism design first and foremost provides the security for the blockchain itself, encouraging miners or proof of stake validators to participate honestly, but more recently it is being applied in prediction markets, "token curated registries" and many other contexts. The nascent RadicalXChange movement has meanwhile spawned experimentation with Harberger taxes, quadratic voting, quadratic financing and more. More recently, there has also been growing interest in using token-based incentives to try to encourage quality posts in social media. However, as development of these systems moves closer from theory to practice, there are a number of challenges that need to be addressed, challenges that I would argue have not yet been adequately confronted.

As a recent example of this move from theory toward deployment, Bihu, a Chinese platform that has recently released a coin-based mechanism for encouraging people to write posts. The basic mechanism (see whitepaper in Chinese here) is that if a user of the platform holds KEY tokens, they have the ability to stake those KEY tokens on articles; every user can make (k)

"upvotes" per day, and the "weight" of each upvote is proportional to the stake of the user making the upvote. Articles with a greater quantity of stake upvoting them appear more prominently, and the author of an article gets a reward of KEY tokens roughly proportional to the quantity of KEY upvoting that article. This is an oversimplification and the actual mechanism has some nonlinearities baked into it, but they are not essential to the basic functioning of the mechanism. KEY has value because it can be used in various ways inside the platform, but particularly a percentage of all ad revenues get used to buy and burn KEY (yay, big thumbs up to them for doing this and not making yet another medium of exchange token!).

This kind of design is far from unique; incentivizing online content creation is something that very many people care about, and there have been many designs of a similar character, as well as some fairly different designs. And in this case this particular platform is already being used significantly:

A few months ago, the Ethereum trading subreddit/r/ethtrader introduced a somewhat similar experimental feature where a token called "donuts" is issued to users that make comments that get upvoted, with a set amount of donuts issued weekly to users in proportion to how many upvotes their comments received. The donuts could be used to buy the right to set the contents of the banner at the top of the subreddit, and could also be used to vote in community polls. However, unlike what happens in the KEY system, here the reward that B receives when A upvotes B is not proportional to A's existing coin supply; instead, each Reddit account has an equal ability to contribute to other Reddit accounts.

These kinds of experiments, attempting to reward quality content creation in a way that goes beyond the known limitations of donations/microtipping, are very valuable; under-compensation of user-generated internet content is a very significant problem in society in general (see "liberal radicalism" and "data as labor"), and it's heartening to see crypto communities attempting to use the power of mechanism design to make inroads on solving it. But unfortunately, these systems are also vulnerable to attack.

Self-voting, plutocracy and

bribes

Here is how one might economically attack the design proposed above. Suppose that some wealthy user acquires some quantity (N)

of tokens, and as a result each of the user's (k)

upvotes gives the recipient a reward of (N \cdot q)

((q)

here probably being a very small number, eg. think (q = 0.000001)

). The user simply upvotes their own sockpuppet accounts, giving themselves the reward of (N \cdot k \cdot q)

. Then, the system simply collapses into each user having an "interest rate" of (k \cdot q)

per period, and the mechanism accomplishes nothing else.

The actual Bihu mechanism seemed to anticipate this, and has some superlinear logic where articles with more KEY upvoting them gain a disproportionately greater reward, seemingly to encourage upvoting popular posts rather than self-upvoting. It's a common pattern among coin voting governance systems to add this kind of superlinearity to prevent self-voting from undermining the entire system; most DPOS schemes have a limited number of delegate slots with zero rewards for anyone who does not get enough votes to join one of the slots, with similar effect. But these schemes invariably introduce two new weaknesses:

- They subsidize plutocracy

, as very wealthy individuals and cartels can still get enough funds to self-upvote.

- They can be circumvented by users bribing

other users to vote for them en masse.

Bribing attacks may sound farfetched (who here has ever accepted a bribe in real life?), but in a mature ecosystem they are much more realistic than they seem. In most contexts where bribing has taken place in the blockchain space, the operators use a euphemistic new name to give the concept a friendly face: it's not a bribe, it's a "staking pool" that "shares dividends". Bribes can even be obfuscated: imagine a cryptocurrency exchange that offers zero fees and spends the effort to make an abnormally good user interface, and does not even try to collect a profit; instead, it uses coins that users deposit to participate in various coin voting systems. There will also inevitably be people that see in-group collusion as just plain normal; see a recent scandal involving EOS DPOS for one example:

Finally, there is the possibility of a "negative bribe", ie. blackmail or coercion, threatening participants with harm unless they act inside the mechanism in a certain way.

In the /r/ethtrader experiment, fear of people coming in and buying

donuts to shift governance polls led to the community deciding to make only locked (ie. untradeable) donuts eligible for use in voting. But there's an even cheaper attack than buying donuts (an attack that can be thought of as a kind of obfuscated bribe): renting

them. If an attacker is already holding ETH, they can use it as collateral on a platform likeCompound to take out a loan of some token, giving you the full right to use that token for whatever purpose including participating in votes, and when they're done they simply send the tokens back to the loan contract to get their collateral back - all without having to endure even a second of price exposure to the token that they just used to swing a coin vote, even if the coin vote mechanism includes a time lockup (as eg. Bihu does). In every case, issues around bribing, and accidentally over-empowering well-connected and wealthy participants, prove surprisingly difficult to avoid.

Identity

Some systems attempt to mitigate the plutocratic aspects of coin voting by making use of an identity system. In the case of the /r/ethtrader donut system, for example, although governance polls

are done via coin vote, the mechanism that determines how many donuts (ie. coins) you get in the first place

is based on Reddit accounts: 1 upvote from 1 Reddit account = (N)

donuts earned. The ideal goal of an identity system is to make it relatively easy for individuals to get one identity, but relatively difficult to get many identities. In the /r/ethtrader donut system, that's Reddit accounts, in the Gitcoin CLR matching gadget, it's Github accounts that are used for the same purpose. But identity, at least the way it has been implemented so far, is a fragile thing....

Oh, are you too lazy to make a big rack of phones? Well maybe you're lookingfor this:

Arguably, attacking these mechanisms by simply controlling thousands of fake identities like a puppetmaster is even easier

than having to go through the trouble of bribing people. And if you think the response is to just increase security to go up to government-level

IDs? Well, if you want to get a few of those you can start exploringhere, but keep in mind that there are specialized criminal organizations that are well ahead of you, and even if all the underground ones are taken down, hostile governments are definitely going to create fake passports by the millions if we're stupid enough to create systems that make that sort of activity profitable. And this doesn't even begin to mention attacks in the opposite direction, identity-issuing institutions attempting to disempower marginalized communities by denying

them identity documents...

Collusion

Given that so many mechanisms seem to fail in such similar ways once multiple identities or even liquid markets get into the picture, one might ask, is there some deep common strand that causes all of these issues? I would argue the answer is yes, and the "common strand" is this: it is much harder, and more likely to be outright impossible, to make mechanisms that maintain desirable properties in a model where participants can collude, than in a model where they can't. Most people likely already have some intuition about this; specific instances of this principle are behind well-established norms and often laws promoting competitive markets and restricting price-fixing cartels, vote buying and selling, and bribery. But the issue is much deeper and more general.

In the version of game theory that focuses on individual choice - that is, the version that assumes that each participant makes decisions independently and that does not allow for the possibility of groups of agents working as one for their mutual benefit, there are mathematical proofs that at least one stable Nash equilibrium must exist in any game, and mechanism designers have a very wide latitude to "engineer" games to achieve specific outcomes. But in the version of game theory that allows for the possibility of coalitions working together, called cooperative game theory

, there are large classes of games that do not have any stable outcome that a coalition cannot profitably deviate from

.

Majority games

, formally described as games of (N)

agents where any subset of more than half of them can capture a fixed reward and split it among themselves, a setup eerily similar to many situations in corporate governance, politics and many other situations in human life, are part of that set of inherently unstable games. That is to say, if there is a situation with some fixed pool of resources and some currently established mechanism for distributing those resources, and it's unavoidably possible for 51% of the participants can conspire to seize control of the resources, no matter what the current configuration is there is always some conspiracy that can emerge that would be profitable for the participants. However, that conspiracy would then in turn be vulnerable to potential new conspiracies, possibly including a combination of previous conspirators and victims... and so on and so forth.

This fact, the instability of majority games under cooperative game theory, is arguably highly underrated as a simplified general mathematical model of why there may well be no "end of history" in politics and no system that proves fully satisfactory; I personally believe it's much more useful than the more famous Arrow's theorem, for example.

There are two ways to get around this issue. The first is to try to restrict ourselves to the class of games that are

"identity-free" and "collusion-safe", so where we do not need to worry about either bribes or identities. The second is to try to attack the identity and collusion resistance problems directly, and actually solve them well enough that we can implement non-collusion-safe games with the richer properties that they offer.

Identity-free and

collusion-safe game design

The class of games that is identity-free and collusion-safe is substantial. Even proof of work is collusion-safe up to the bound of a single actor having ~23.21% of total hashpower, and this bound can be increased up to 50% withclever engineering. Competitive markets are reasonably collusion-safe up until a relatively high bound, which is easily reached in some cases but in other cases is not.

In the case of governance

and content curation

(both of which are really just special cases of the general problem of identifying public goods and public bads) a major class of mechanism that works well is futarchy

- typically portrayed as "governance by prediction market", though I would also argue that the use of security deposits is fundamentally in the same class of technique. The way futarchy mechanisms, in their most general form, work is that they make "voting" not just an expression of opinion, but also a prediction

, with a reward for making predictions that are true and a penalty for making predictions that are false. For example,my proposal for "prediction markets for content curation DAOs" suggests a semi-centralized design where anyone can upvote or downvote submitted content, with content that is upvoted more being more visible, where there is also a "moderation panel" that makes final decisions. For each post, there is a small probability (proportional to the total volume of upvotes+downvotes on that post) that the moderation panel will be called on to make a final decision on the post. If the moderation panel approves a post, everyone who upvoted it is rewarded and everyone who downvoted it is penalized, and if the moderation panel disapproves a post the reverse happens; this mechanism encourages participants to make upvotes and downvotes that try to "predict" the moderation panel's judgements.

Another possible example of futarchy is a governance system for a project with a token, where anyone who votes for a decision is obligated to purchase some quantity of tokens at the price at the time the vote begins if the vote wins; this ensures that voting on a bad decision is costly, and in the limit if a bad decision wins a vote everyone who approved the decision must essentially buy out everyone else in the project. This ensures that an individual vote for a "wrong" decision can be very costly for the voter, precluding the possibility of cheap bribe attacks.

However, that range of things that mechanisms of this type can do is limited. In the case of the content curation example above, we're not really solving governance, we're just scaling

the functionality of a governance gadget that is already assumed to be trusted. One could try to replace the moderation

panel with a prediction market on the price of a token representing the right to purchase advertising space, but in practice prices are too noisy an indicator to make this viable for anything but a very small number of very large decisions. And often the value that we're trying to maximize is explicitly something other than maximum value of a coin.

Let's take a more explicit look at why, in the more general case where we can't easily determine the value of a governance decision via its impact on the price of a token, good mechanisms for identifying public goods and bads unfortunately cannot be identity-free or collusion-safe. If one tries to preserve the property of a game being identity-free, building a system where identities don't matter and only coins do, there is an impossible tradeoff between either failing to incentivize legitimate public goods or over-subsidizing plutocracy

.

The argument is as follows. Suppose that there is some author that is producing a public good (eg. a series of blog posts) that provides value to each member of a community of 10000 people. Suppose there exists some mechanism where members of the community can take an action that causes the author to receive a gain of $1. Unless the community members are extremely

altruistic, for the mechanism to work the cost of taking this action must be much lower than $1, as otherwise the portion of the benefit captured by the member of the community supporting the author would be much smaller than the cost of supporting the author, and so the system collapses into a tragedy of the commons where no one supports the author. Hence, there must exist a way to cause the author to earn $1 at a cost much less than $1. But now suppose that there is also a fake community, which consists of 10000 fake sockpuppet accounts of the same wealthy attacker. This community takes all of the same actions as the real community, except instead of supporting the author, they support another

fake account which is also a sockpuppet of the attacker. If it was possible for a member of the "real community" to give the author $1 at a personal cost of much less than $1, it's possible for the attacker to give themselves

$1 at a cost much less than $1 over and over again, and thereby drain the system's funding. Any mechanism that can help genuinely under-coordinated parties coordinate will, without the right safeguards, also help already coordinated parties (such as many accounts controlled by the same person) over-coordinate

, extracting money from the system.

A similar challenge arises when the goal is not funding, but rather determining what content should be most visible. What content do you think would get more dollar value supporting it: a legitimately high quality blog article benefiting thousands of people but benefiting each individual person relatively slightly, or this?

Or perhaps this?

Those who have been following recent politics "in the real world" might also point out a different kind of content that benefits highly centralized actors: social media manipulation by hostile governments. Ultimately, both centralized systems and decentralized systems are facing the same fundamental problem, which is that the "marketplace of ideas" (and of public goods more generally) is very far from an "efficient market" in the sense that economists normally use the term

, and this leads to both underproduction of public goods even in "peacetime" but also vulnerability to active attacks. It's just a hard problem.

This is also why coin-based voting systems (like Bihu's) have one major genuine advantage over identity-based systems (like the Gitcoin CLR or the /r/ethtrader donut experiment): at least there is no benefit to buying accounts en masse, because everything you do is proportional to how many coins you have, regardless of how many accounts the coins are split between. However, mechanisms that do not rely on any model of identity and only rely on coins fundamentally cannot solve the problem of concentrated interests outcompeting dispersed communities trying to support public goods; an identity-free mechanism that empowers distributed communities cannot avoid over-empowering centralized plutocrats pretending to be distributed communities.

But it's not just identity issues that public goods games are vulnerable too; it's also bribes. To see why, consider again the example above, but where instead of the "fake community" being 10001 sockpuppets of the attacker, the attacker only has one identity, the account receiving funding, and the other 10000 accounts are real users - but users that receive a bribe of $0.01 each to take the action that would cause the attacker to gain an additional $1. As mentioned above, these bribes can be highly obfuscated, even through third-party custodial services that vote on a user's behalf in exchange for convenience, and in the case of "coin vote" designs an obfuscated bribe is even easier: one can do it by renting coins on the market and using them to participate in votes. Hence, while some kinds of games, particularly prediction market or security deposit based games, can be made collusion-safe and identity-free, generalized public goods funding seems to be a class of problem where collusion-safe and identity-free approaches unfortunately just cannot be made to work.

Collusion resistance and

identity

The other alternative is attacking the identity problem head-on. As mentioned above, simply going up to higher-security centralized identity systems, like passports and other government IDs, will not work at scale; in a sufficiently incentivized

context, they are very insecure and vulnerable to the issuing governments themselves! Rather, the kind of "identity" we are talking about here is some kind of robust multifactorial set of claims that an actor identified by some set of messages actually is a unique individual. A very early proto-model of this kind of networked identity is arguably social recovery in HTC's blockchain phone:

The basic idea is that your private key is secret-shared between up to five trusted contacts, in such a way that mathematically ensures that three of them can recover the original key, but two or fewer can't. This qualifies as an "identity system" - it's your five friends determining whether or not someone trying to recover your account actually is you. However, it's a special-purpose identity system trying to solve a problem - personal account security - that is different from (and easier than!) the problem of attempting to identify unique humans. That said, the general model of individuals making claims about each other can quite possibly be bootstrapped into some kind of more robust identity model. These systems could be augmented if desired using the "futarchy" mechanic described above: if someone makes a claim that someone is a unique human, and someone else disagrees, and both sides are willing to put down a bond to litigate the issue, the system can call together a judgement panel to determine who is right.

But we also want another crucially important property: we want an identity that you cannot credibly rent or sell. Obviously, we can't prevent people from making a deal "you send me $50, I'll send you my key", but what we can

try to do is prevent such deals from being credible

- make it so that the seller can easily cheat the buyer and give the buyer a key that doesn't actually work. One way to do this is to make a mechanism by which the owner of a key can send a transaction that revokes the key and replaces it with another key of the owner's choice, all in a way that cannot be proven. Perhaps the simplest way to get around this is to either use a trusted party that runs the computation and only publishes results (along with zero knowledge proofs proving the results, so the trusted party is trusted only for privacy, not integrity), or decentralize the same functionality through multi-party computation. Such approaches will not solve collusion completely; a group of friends could still come together and sit on the same couch and coordinate votes, but they will at least reduce it to a manageable extent that will not lead to these systems outright failing.

There is a further problem: initial distribution of the key. What happens if a user creates their identity inside a third-party custodial service that then stores the private key and uses it to clandestinely make votes on things? This would be an implicit bribe, the user's voting power in exchange for providing to the user a convenient service, and what's more, if the system is secure in that it successfully prevents bribes by making votes unprovable, clandestine voting by third-party hosts would also

be undetectable. The only approach that gets around this problem seems to be.... in-person verification. For example, one could have an ecosystem of "issuers" where each issuer issues smart cards with private keys, which the user can immediately download onto their smartphone and send a message to replace the key with a different key that they do not reveal to anyone. These issuers could be meetups and conferences, or potentially individuals that have already been deemed by some voting mechanic to be trustworthy.

Building out the infrastructure for making collusion-resistant mechanisms possible, including robust decentralized identity systems, is a difficult challenge, but if we want to unlock the potential of such mechanisms, it seems unavoidable that we have to do our best to try. It is true that the current computer-security dogma around, for example, introducing online voting is simply "don't", but if we want to expand the role of voting-like mechanisms, including more advanced forms such as quadratic voting and quadratic finance, to more roles, we have no choice but to confront the challenge head-on, try really hard, and hopefully succeed at making something secure enough, for at least some use cases.