Forever Blowing Bitcoins: Social Structure and Speculative Bubbles in Cryptocurrencies

Nicolas Della Penna [*]
ANU
n@nikete.com

Eaman Jahani [†], Peter Krafft
MIT
eaman,pkrafft@mit.edu

Harper Reed
harper@nata2.org

Octavio Bunge
Universidad de Belgrano
octavio.bunge@comunidad.ub.edu.ar

Sandy Pentland
MIT Media Lab
sandy@media.mit.edu

Julian McAuley
UC San Diego
cpalmer@prl.com

ABSTRACT

We study the power of structural features of the social network around cryptocurrencies to understand the severity of bubbles that occur in them. Our goal is to see whether bubbles in cryptocurrency markets can be predicted purely from structural features derived from the social networks that preceed them. All our measures are constructed on the social network before the relevant cryptocurrency is ever traded.

1. INTRODUCTION

Speculative bubbles, since at least the early 18th century South Sea Bubble are perceived [1] to periodically take over markets. The public notoriety of Bitcoin and the massive price increases and their associated publicity lead to an explosion of attempts to create "the next bitcoin" often referred to as "cryptocurrencies" or "coins", and a vibrant set of exchanges where these are traded, either for each other or money. The majority of these coins have no viable uses, and their markets would appear driven largely by speculation. Many of them appear to be nothing but attempts at turning a quick profit from inflating the implied valuation of a coin shortly after creating it. This is driven by the extremely low cost and effort required to create a new coin, with most being minimal changes to parameters and branding of a pre-existing codebase. Those who make and trade these coins communicate largely online, and much of their activity is concentrated on public forums, price and volume data from their exchanges is freely available and widely aggregated, and the source code to all coins is public. This makes cryptocoins an almost ideal window in the social life of a market mania [?]. Such study can serve in the computational social sciences a role analogous to that of lesion studies do in neuropsychology.

We present a novel dataset that combines measures derived from the social network in an online forum, market data aggregated over dozens of exchanges, and properties of software implementing hundreds of cryptocoins. In the forum we identify the introducers of each coin and build measures of their position in the network based on which users have engaged with them threads in the forum before the coin is announced. We dentify 376 coins that are announced by users of the forum and which can be mapped to price and volume data from exchanges. From the price and transaction volume data we build measures of the subsequent activity that results from trading in the coin. We also asses if coins posibly embody technological inovation based on having more than trivial modifications to previously existing coins sourcecode.

While the mechanisms that drive bubbles have been theoretically [?, ?, ?, ?] and experimentally [?] in the lab, an exaustive dataset on the social network of those promoting the asset has not been previously available. While the magnitude of the assets traded is small relative to most financial and commodity markets, it is much larger than even the most lavishly funded experimenter could hope for. The largest bubble in our dataset, AuroraCoin, reaches a valuation of 1 billon USD on March , with reported daily trading volumes of 6.8M USD, and sheds 90% of its values in a week, and 99% of its value in well under a year. For context, this is equivalent to 1/4 of Icelands entire foreigh exchange reserves in 2014, [2], the population of which AuroraCoin promoters claimed they would distribute half of the coins to.

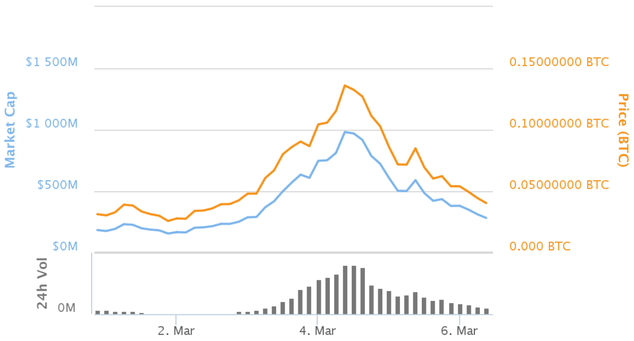
Using the price and volume data we construct measures of both the magnitude (how many dollars worth of trading happened in the asset) and the severity of bubbles (which we define as for each dollar invested at the peak what could be recovered if selling at the volume weighted average prices

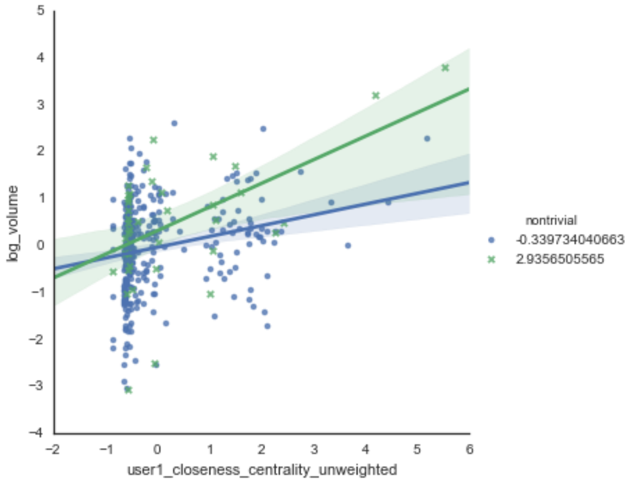
[*] is a first co-author of this paper.

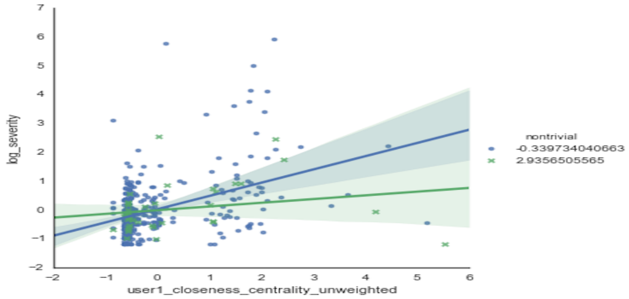
[†] is a first co-author of this paper.

[1] [?] pp 127-31 for references, a particularly is [?] a case study of a well-informed investor in the South Sea bubble that invested knowingly in the bubble, and found that it was profitable.

[2] 4.1 Billon USD , The World Bank, Global Economic Monitor, accessed October 2015







observed since). By considering the community structure that exists in the forum before a coin is introduced we are able to predict part of the variation in both the severity and magnitude of the resulting bubble: our best model explains 10% of the variation in the severity and of the magnitudein our of sample using a penalized linear model and measuring performance out of sample using cross validation. The main driver of our explanatory power is the centrality of a user in the directed network derived from the forum. Both the severity and the magnitude of bubbles increases with the centrality of the user who introduces the coins in the forums. Interestingly this effect is concentrated in different ways depending on weather the coin software is more than a trivial modification: trivial coins have more severe bubbles the more central their introducers are, while volume is greater the more central the introducer of a nontrivial coin is.

While states can create the demand required for a currency system to run by compelling tax payment in it (for a recent example), non state sponsored currencies must find some other ways of creating demand. The initial market for which bitcoin has been used (prices denominated in it, transactions only in it) was drug sales.[3] Since the cost of producing a new coin is effectively zero, new currencies have thus been floated with every single drug name possible. Many chains can claim to the same name, so exchanges with volume (since speculation is the only possible use of almost all of the coins) become de-facto arbitrators of who has a minimally viable claim.[4]

2. LITERATURE

This work is at the intersection of three literatures: in economics and finance on the study of speculative bubbles, in network science on the prediction of outcomes based on features of an individual in a network, and a in computer science, largely centered on the security comunity, studying cryptocurrencies.

2.1 Bubbles

Perhaps the most striking line of research on bubbles in economics with respect to cryptocurrencies is the study of markets where the asset is worthless and this is common knowledge. Recently [**?**] studies both theoretically and experimentally in the laboratory such a bubble. The driving force is that some traders "do not know where they stand in the market sequence, the game allows for a bubble at the Nash equilibrium when there is no cap on the maximum price.". In the context of cryptocurrencies the lack of knowledge around the sequence position maps to uncertainty about ones place in the technology adoption knowledge and adoption curve, while the dificulty in upper bounding the potential market value of cryptocurrencies provides the lack of cap on the maximum price.

A large literature in finance empirically examines herding by financial analysts, for a recent example @articlejegadeesh2009analysts, title=Do analysts herd? An analysis of recommendations and market reactions, author=Jegadeesh, Narasimhan and Kim, Woojin, journal=Review of Financial Studies, pages=hhp093, year=2009, publisher=Soc Financial Studies It also looks at the properties of analysts who disagre with their peers and of their forecasts

[**?**] weather analysts are likely to disagree with their peers-for a prominent recent example classifies analysts' earnings forecasts as herding or bold and finds that (1) boldness likelihood increases with the analyst's prior accuracy, brokerage size, and experience and declines with the number of industries the analyst follows, consistent with theory linking boldness with career concerns and ability; (2) bold forecasts are more accurate than herding forecasts; and (3) herding forecast revisions are more strongly associated with analysts' earnings forecast errors (actual earningsâĂŤforecast) than

[3] A overview of the different drug marketplaces and estimated transaction volumes can be found in [**?**]. To the best of the authors knowledge no other sector beyond speculation has even remotely substantial volume at present; a very primitive form of unregulated gambling Satoshi Dice, did for a brief pointing the past)

[4] While it is theoretically possible to engage in a distributed protocol to exchange between two cryptocurrencies, see part II of lecture 10 in [**?**]

are bold forecast revisions. Thus, bold forecasts incorporate analysts' private information more completely and provide more relevant information to investors than herding forecasts."

Being able to observe the direction of attention in the network is a key characteristic of our dataet, and provids us greater range of network measures that can be constructed. In particular since information differentials between nodes in a network are often systematic, the directionality of edges, of whom is paying attention to whom, matters. Analysts who cover stocks can be considered in anundirected weighted networks based on what proportion of stocks they cover in common, this is used by [?] to build an indicator based on the average degree of nodes and the average weighted clustering coefficint. They find herding accross all industries in various degrees, and that there is industry level variation on wether it is informed hearding in reaction to public news or uninformed speculation.

[?] examine whether access to management at broker-hosted investor conferences leads to more informative research by analysts. We find analyst recommendation changes have larger immediate price impacts when the analystŒşs firm has a conference-hosting relation with the company. The effect increases with hosting frequency and is strongest in the days following the conference. Conference-hosting brokers also issue more informative, accurate, and timely earnings forecasts than non-hosts. Our findings suggest that access to management remains an important source of analystsŒş informational advantage in the post-Regulation Fair Disclosure world.

[?] This study shows how the emotional phases that accompany market crisis can be related to an underlying cycle of actions, attributions, and regulatory reactions among participants in the market environment. The action-attribution-regulation process is here called ÂŞenactmentÂŤ, in order to focus on how market participants create the environment which then impinges on their activity. This process is then illustrated with a case study of the 1980 crisis in the silver futures market, when prices soared from 10 per ounce to 50 per ounce and fell back to 10 per ounce in seven months. The traditional mania/distress/panic model of speculative bubbles is reframed as a cycle of organising, focusing on the strategic actions of buyers, sellers, bankers, and government agencies. The paper shows how the crisis, enacted by market participants who created speculative opportunities, was resolved through the cooperation of powerful organisations that sought to protect the solvency of insiders and the integrity of the market. This view of market process suggests a cycle of action and institutional constraint which shapes the structure of market environments.

2.2 Prediction from networks

"The Structural Virality of Online Diffusion" (Management Science)

"Here we propose a formal measure of what we label âĂIJstructural viralityâĂİ that interpolates between two conceptual extremes: content that gains its popularity through a single, large broadcast, and that which grows through multiple generations with any one individual directly responsible for only a fraction of the total adoption"

"We find that across all domains and all sizes of events, online diffusion is characterized by surprising structural diversity. Popular events, that is, regularly grow via both broadcast and viral mechanisms, as well as essentially all conceivable combinations of the two."

"we find that the correlation between the size of an event and its structural virality is surprisingly low, meaning that knowing how popular a piece of content is tells one little about how it spread"

"We find that while several of our empirical findings are consistent with such a model, it does not replicate the observed diversity of structural virality"

Network Diversity and Economic Development REPORT Network Diversity and Economic Development Nathan Eagle

Social networks form the backbone of social and economic life. Until recently, however, data have not been available to study the social impact of a national network structure. To that end, we combined the most complete record of a national communication network with national census data on the socioeconomic well-being of communities. These data make possible a population-level investigation of the relation between the structure of social networks and access to socioeconomic opportunity. We find that the diversity of individualsâĂŹ relationships is strongly correlated with the economic development of communities.

"Hence, highly clustered, or insular, social ties are predicted to limit access to social and economic prospects from outside the social group, whereas heterogeneous social ties may generate these opportunities from a range of diverse contacts (1, 2)."

"Although both social and spatial network diversity scores were strongly correlated with IMD rank, we found a weaker positive correlation present using number of contacts and a negative correlation for communication volume."

"For example, whereas inhabitants of Stoke-on-Trent, one of the least prosperous regions in the UK, averaged a higher monthly call volume than the national average, they have one of the lowest diversity scores in the country. Similarly prosperous Stratford-upon-Avon has inhabitants with extremely diverse networks, despite no more communication than the national average. "

Predicting Spending Behavior Using Socio-mobile Features: free version:

Social behavior can be used to predict spending behavior in couples in regards to their prepensity to diversify the businesses they explore, become loyal customers and overspend. The results show that mobile phone social interaction patters can be more predictive than personality based features when predicting spending behavior.

"We find that social behavior measured via face-to-face interaction, call, and SMS logs, can be used to predict the spending behavior for couples in terms of their propensity to explore diverse businesses, become loyal customers, and overspend"

"results show that mobile phone based social interaction patterns can provide more predictive power on spending behavior than personality based features. Interestingly, we find that more social couples also tend to overspend."

Money Walks: Implicit Mobility Behavior and Financial Well-Being:

Spatiotemporal traits such as exploration, engagement and elasticity can be used to predict future finanical difficulties.

"Hence, in this work we study a large-scale cross-sectional dataset of human spending across space and time, and connect it to the biological phenomena of âĂIJforaging,âĂİ a basic pattern of animal movement to gather foods and resources."

"we analyzed a corpus of hundreds of thousands of human economic transactions and found that financial outcomes for individuals are intricately linked with their spatiotemporal traits like exploration, engagement, and elasticity. Such features yield models that are 30% to 49% better at predicting future financial difficulties than the comparable demographic models."

"As shown in Fig 2, individuals with lower levels of education (High School, Middle School, or Primary School) were found to be more likely to be late for their payments and get into financial trouble. Users with higher age were marginally less likely to overspend, miss payments, or get into financial trouble. Last, male customers and married customers were less likely to miss their payments."

"The figure also shows that multiple mobility behavior features were statistically correlated with outcome variables, even after controlling for the effect of abovementioned demographic variables of age, gender, marital status, education, and work type."

"the behavioral features were found to be more significantly associated (in terms of p-values) and contain higher predictive power (in terms of odds ratios being further away from 1.0 in either direction) as compared to the demographic features."

"The evidence so far indicating that each of the spatiotemporal behavioral descriptors has significant association with different financial outcomes motivates their combination to predict the financial outcome"

Predicting personality using novel mobile phone-based metrics free version:

"Using a set of novel psychology-informed indicators that can be computed from data available to all carriers, we were able to predict usersâĂŹ personality with a mean accuracy across traits of 42

"The goal of the present research is to show that usersâĂŹ personalities can be reliably inferred from basic information accessible from all mobile phones and to all service providers."

"The model predicted whether phone users were low, average, or high in neuroticism, extraversion, conscientiousness, agreeableness, and openness with an accuracy of 54

2.3 Bitcoin and Cryptocurrencies

heuristic clustering to group Bitcoin wallets based on evidence of shared authority, and then using re-identification attacks (i.e., empirical purchasing of goods and services) to classify the operators of those clusters. From this analysis, we characterize longitudinal changes in the Bitcoin market, the stresses these changes are placing on the system, and the challenges for those seeking to use Bitcoin for criminal or fraudulent purposes at scale." [?]

fistful of bitcoins Bitcoin is a purely online virtual currency, unbacked by either physical commodities or sovereign obligation; instead, it relies on a combination of cryptographic protection and a peer-to-peer protocol for witnessing settlements. Consequently, Bitcoin has the unintuitive property that while the ownership of money is implicitly anonymous, its flow is globally visible. In this paper we explore this unique characteristic further, using heuristic clustering to group Bitcoin wallets based on evidence of shared authority, and then using re-identification attacks (i.e., empirical purchasing of goods and services) to classify the operators of those clusters. From this analysis, we characterize longitudinal changes in the Bitcoin market, the stresses these changes are placing on the system, and the challenges for those seeking to use Bitcoin for criminal or fraudulent purposes at scale.

[?] Measuring the longitudinal evolution of the online anonymous marketplace ecosystem

[?] How Did Dread Pirate Roberts Acquire and Protect His Bitcoin Wealth

3. DATA DESCRIPTION

3.1 Prices, Exchanges, and Coin charactertistics

Our main outcome measures are the severity of drop in the value of a unit of the asset, and the magnitude in USD of the transactions in them. To obtain data for them we scrape three market aggregators

We operationalize the intensity of a bubble as the proportion of a 1 dollar that would be lost buying at the maximum price and selling after that proportionally to the volume of the market till the present, we call this severity. We define the volume as the sum of the contemporaneous dollar (todo check) volume of trade.

As a secondary outcome measure we consider the number of exchanges that list the coin.

3.2 Forum Discussions

In order to study the effect of communication network around cryptocoins on price variations, we collected all the posts from the most popular cryptocurrency online community, bitcointalk. Our data consisted of all the posts that were made between January 2010 and July 2015 on the most active crypto-related forums:

1. **Bitcoin Discussion:** This is the oldest forum on the website which mainly focuses on issues only related to Bitcoin. Interestingly, Satoshi Nakamoto, the alleged creator of Bitcoin made the first post on this forum in January 2010 and was active until January 2011. The presence of Satoshi in the data set enables us to study the position of various actors in the online community relative to Satoshi and its relation with the success or failure of cryptocoins they advocate or reject.
2. **Altcoin Discussion:** This is the most active forum in the community with more than 730,000 posts as of July 2015, and dating back to June 2011. The discussions in this forum mainly evolve around alternative currencies other than Bitcoin. Users often discuss the merits or flaws of various altcoins or simply exchange technical information.
3. **Announcement (Altcoin):** Community announcements such as development of exchange client or addition of new features are made here. This is an important forum in our study as the creation of new altcoins are announced here. Whenever a new altcoin is announced to the community, the announcement is tagged with string ANN. This enables us to detect announcement of new coins into the market and identify the users who introduced them for the first time.
4. **Mining (Altcoin):** Technical issues pertaining to mining (i.e. validating transactions) altcoins are discussed here.
5. **Marketplace (Altcoin):** This forum contains the discussions on a wide-range of market-related issues, such as price or volume trends, possible pump and dump schemes and exchange tips.

TODO:Do we want any other descriptive stats on the forum other than those mentioned here?

Each forum consists of many subjects or threads initiated by different users. Each thread contains several posts or replies, with an average of 10 posts per thread. The reply structure within each thread constitutes the basis of our forum network, discussed below. Each post has several fields which contain valuable information in our context.

1. **Subject:** Usually, the initiator of the thread chooses subject and all the following posts inherit the same subject.
2. **Content:** The actual text of the post.
3. **Position in the thread**: The later posts in the thread might not be as important as earlier posts and could be about issues other than the original topic of the thread.
4. **Author**
5. **Date**

The community had only 10000 unique users until early 2013, however it grew considerably faster after 2013 and reached about 70000 by early 2015. Nevertheless, there are only 10000 active users within any 30 day period on average.

4. FORUM INTERACTION NETWORK

Given the forum discussion data at the level of individual posts, we construct a network capturing the discussion patterns among users. The structural properties of this network form the basis of our analysis on a per-coin basis. In this network, nodes are the forum users and *directed edges* point from posters within each thread to thread-initiators. The omission of edges based on simple co-appearance within a thread leads to a sparser network which isolates the communication patterns around “dialogue-shapers”. The edges in the discussion network are weighted by the number of times a poster replies to a thread-initiator in different threads (i.e. multiple replies by the same user within the same thread count only as once). In this context, edge weights capture the level of engagement thread initiators receive from the community and the amount of information a poster receives from thread initiators. Furthermore, our network construction method uses all the interactions since the inception of bitcointalk in creating new edges or updating their weights. The unlimited retention of any such (replier to thread initiator) interaction captures relevant information on seniority and community influence which are obtained through long-term and persistent presence in the forums.

Prior to construction of the network, we merged posts from all forums into a single large forum since the community base of all five forums mentioned above is made of the same users and we are mostly concerned about influence and aggregate information flow among users, rather than the exact topic of the discussion. The network construction involves replaying all the posts over time sorted by their date and updating the discussion network accordingly. Whenever a new altcoin is introduced in the forum for the first time, the user who introduced it and a snapshot of the network is taken. We analyze the discussion network only up to the first time each coin is introduced to the community, in order to avoid any possible confounding between a coin’s price movement and the extra attention it receives in the community due the same price changes. Our method uses the position of the first introducer in the network snapshot and the general structure of her neighborhood for extracting various network measures corresponding to that coin. Our final analysis examines these per-coin measures for evaluating the performance of each coin.

The majority of such introductions are made in the *Announcement* forum and are preceded with the “ANN” tag. We look for the first mention of both the coin symbol **and** its descriptive name in the subject of a thread which contains the announcement tag. The first mentions of either the coin symbol **or** the its name are used fall-back in case the more restrictive **and** requirement did not detect the coin. Using this method, we were able to detect the first introduction of 554 altcoins out of 679. The forum user who initiated such a thread is assigned as the introducer of the coin to the community.

TODO: add validation results table wrt mapofcoins data here

4.1 Nontrivial coins

Many of the coins available in the exchanges are trivial modifications of another coin in that they only change parameters such as the name, the number of total mineable coins, or the transaction time between blocks. These coins production cost is virtualy zero[5].

To attempt to capture this we analize data from mapofcoins.com which includes a genealogy of coins and data from the github page of coins not available on maofcoins.com. If the coin to be analyzed has a parent and the algorithm it uses differs from the parent or if it has no parent, it is labeled as nontrivial, meaning that the coin implemented something that did not previously exist, and is not just a fork with only parameters changed, such as total mineable coins, transaction speed, etc.

5. ANALYSIS VARIABLES

5.1 Prices and Exchanges

Our main outcome measures are the severity of the inflation an asset price, and the magnitude of money transacted in it. We operationalize the intensity of a bubble as the proportion of a 1 dollar (TODO check currency base) that would be lost buying at the maximum price and selling after that proportionally to the volume of the market till the present, we call this severity. We define the volume as the sum of the contemporaneous dollar (todo check) volume of trade. As a secondary outcome measure we consider the number of exchanges that list the coin.

5.2 Network Structure

In this section, we discuss the various metrics extracted from discussion networks and used as independent variables in the regression analysis. Many of these variables are standard metrics in graph theory designed to capture node centrality is specific scenarios [?]. As mentioned before, each coin is associated with a forum user and a discussion network which corresponds to the state of the forum at the time the user introduced the coin to the community. All of our node-level variables refer to the user introducing the coin. Below, we list the network variables included in the analysis. We used Python igraph implementation for computing the network-related metrics [?].

1. **Introducer number of posts:** The total number of posts (thread-initiations or simple replies) the coin introducer has made at the time she introduces the coin. It captures the user's level of activity in the community.
2. **Introducer number of threads:** The total number of threads the coin introducer has made. Users who start many threads are more likely to receive incoming edges and to shape the dialogue in the community.
3. **Seniority:** It is the number of days since the user's first post in the forums. We use this as a proxy for user's seniority in the community.
4. **Incoming degree:** The (incoming) degree centrality captures the role of dialogue-shapers in the community as it is the number of unique users who have replied to any of the focal user's threads.
5. **Outgoing degree:** The (outgoing) degree centrality captures the role followers in the community as it is the number of unique thread initiators the focal user has ever replied to.
6. **Total degree:** The (undirected) degree centrality captures total level of user's involvement in the community in any of the two forms above.
7. **Clustering Coefficient:** A measure embeddedness or triadic closure, it is the fraction of focal user's triads that are closed. It does not use the direction on the edges and measures how tightly knit the focal user is connected to the other users who have ever engaged with her either by replying to her thread or receiving a reply from her. The triadic closure is closely related to the the principle of balance which states that if two user pairs A-B and B-C are connected, the existence of a tie between A and C on the triad further strenghtens it and removes any potential strain that could exist between A-B and B-C relation. In general, ideas are more likely to be reinforced and persistent in a triad if it is closed. Such a positive effect of balance or triadic closure on tie qualities and their persistence is shown to exist in online social network such as Twitter [?], and we believe the same argument applies to this scenario. TODO: DO WE NEED A BETTER INTERPRETATION HERE FOR TRIADIC CLOSURE?
8. **Unweighted closeness centrality:** While degree centrality measures the level of user engagement in the community, it only examines the local structure around the user. In other words, it's possible for the focal user to have a high degree centrality, but only in an isolated subcommunity. Closeness, while closely related to degree centrality, measures the level of user's engagement with the global network *either directly or indirectly.* It is relevant in many scenarios, including the online discussions, as information spreads through the shortest paths. Closeness centrality for the ith user is defined as:

$$C_i = \sum_{j=1}^{N} \frac{N-1}{|S_{ij}|} \tag{1}$$

where S_{ij} denotes the set of users on the shortest path from i to j. It is a normalized sum of distances from the focal user (i) to all other users (j), where edges all are weighted with a distance of 1. The sum of all inverse distances is normalized by the number of users present in the network at the time of coin introduction, so that the comparsion between the closeness centrality of various users (who introduce the coins) at different times is valid.

In our context, a user with high incoming closeness centrality has initiated many threads and recieved replies from a diverse set of users who themselves are close to a large set of diverse users. Similarly, a user with high outgoing closeness centrality has replied to a diverse set of thread-initiators who themselves are close to a large set of diverse users. Our analysis consisted of three versions of the unweighted closeness centrality:

[5] there is a cottage industry that offers the creation of binaries and provisioning of mining and bandwidth as a bundled service that require no technical skill to create form the user, examples are Coingen or Coincreator

(a) **Incoming:** Only the directed paths leading to the focal user are used. In other words, it measures closeness of the whole network to the focal user. Users who start many threads are likely to have higher incoming closeness centrality.

(b) **Outgoing:** Only the directed paths starting from the focal user to all the other users are used. In other words, it measures closeness of the focal user to the whole network. Users who reply to many threads are likely to have higher outgoing closeness centrality.

(c) **Undirected:** The paths both from and to the focal users are used. Users who inititate and reply to many threads are likely to have higher undirected closeness centrality.

9. **Weighted closeness centrality:** Similar to the unweighted closeness centrality above, we computed three different versions, with the exception that edges are weighted to indicate the distance or level of ineraction intensity between the two users. The edges weights in our discussion network are determined by the frequency of interactions between two users; and as two users interact more, they are deemed to be closer in their shortest path. Thus in the computation of weighted closeness centralities, we use the reciprocal of the weights as the distance between two users.

$$C_i = \sum_{j=1}^{N} (N-1) \sum_{e \in S_{ij}} w_e \qquad (2)$$

where e denotes an edge in S_{ij} the set of users on the shortest path from i to j. w_e is the weight of edge e determined by the number of interactions between the end points.

10. **betweenness centrality weighted:**
11. **satoshi distance:**
12. **satoshi pagerank weighted:**
13. **pagerank weighted:**

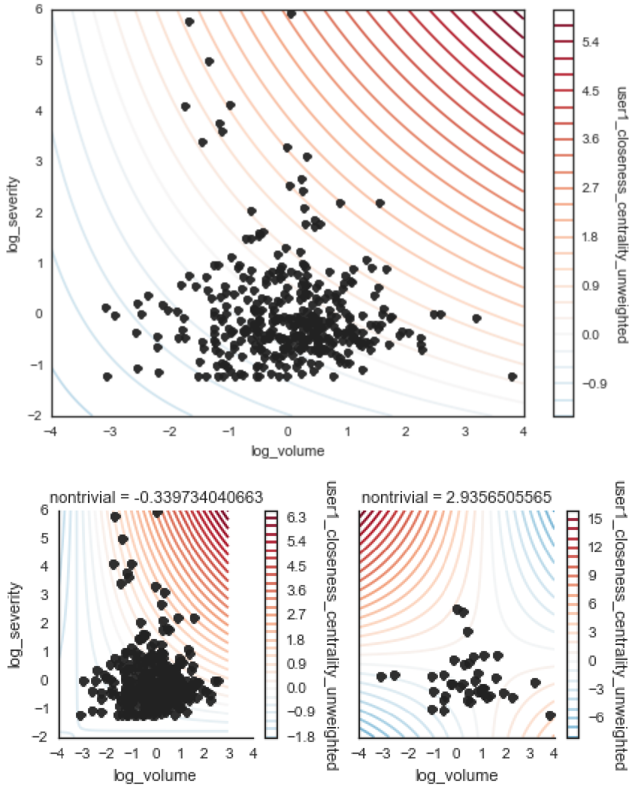
6. METHODS

Initially we we start with a baseline model that considers only the user characteristics that are easily observable from their activity on the forum before the anouncement: the number of posts and of subjects, the time since they first post, the number of users that they have responded to and received responses from. These network measures are possible for any generic discussion, we introduce two further sets of variables to enrich our models that rely on domain knowledge of the underlying assets: satoshi network measures, and weather a given coin is embodied in new software or if it is simply a change in name and parameters of the codebase used by a different coin.

We estimate linear regularized least squares (ElasticNet cite TODO) using a combination of L1 and L2 norm, with their parameters set by 5 fold cross validation. We then estimate a OLS model of the support of the variables and calculate White robust standard errors, to allow for model introspection. Disclaimer that the regularization might make them not match (TODO: add set with normal SE that is estimated with the regularization, in results compare the coefficients) To evaluate nonlinearities and interactions in the model we fit a gradient boosted machine on the full support, cross validating its hyper parameters; as well as on the OLS selected subset. TODO add graphs showing interactions and nonlinearities; table with model comparisons.

The initial analysis pipeline and debugging, hyperparameter setting was done using only th initial 270 of the eventual 560 in the sample. The full set of samples used for these estimates was only estimated before writing the results section. The method will not be revised beyond this point.

7. RESULTS



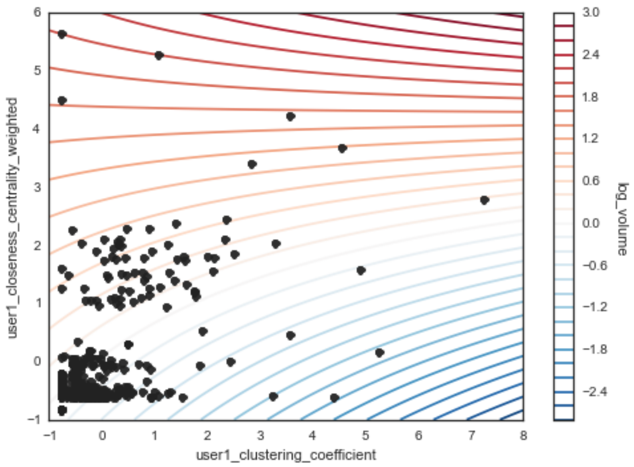
7.1 Limitations & Future Work

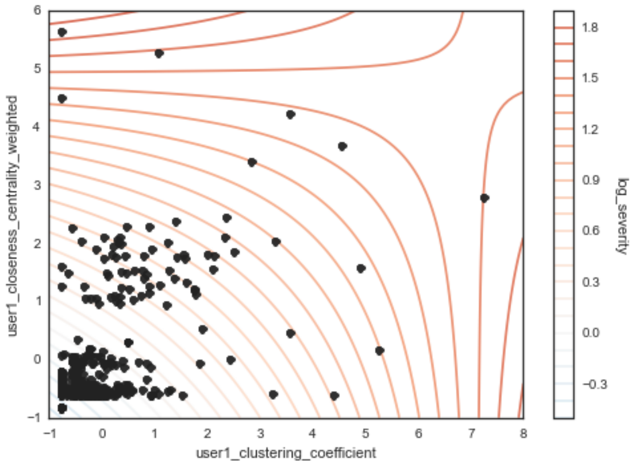
The features of the node in the graph we use as well as the construction ofthe graph, while informed by the literature and theory, could be substantially imporoved. We could attempt to learn the features from a more raw version of the forums, or at least learnthe parametrization of our constructor, or some larger space of weighted, or potentially labeled edges with the language used. The cross sectional design with time separation does not allow us to take advantage of intra-coin variaiton.

Beyond the forums the code repositories could be epxloited in futre work as a rich source of variation.

8. CONCLUSION

The total variance accounted for is small, so you need a discussion like ”results suggest that bubble dynamics may be strongly influenced by a core set of participants, but that traditional network measures on the aggregate discussion graph do not provide a tight characterization of this





core group. We are looking at coarse (positive/negative, detailed/cursory) semantic analysis of the discussion, and evidence of prior cooperation between pairs of participants in other altcoin markets, in order to attempt this more accurate characterization of core participants and their actions.

9. ACKNOWLEDGMENTS

APPENDIX

Table 1: Log(Volume) model OLS parameter estimates with heteroskedasticity robust SE after ElasticNet variable selection

	Activity	Nontrivial	Satoshi	Network	Weighted	Network*Nontrivial	All
Intercept	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
nontrivial		0.15*** (0.05)	0.13** (0.05)	0.06 (0.05)	0.06 (0.05)	0.04 (0.05)	0.10** (0.05)
closeness centrality unweighted				0.24*** (0.06)		0.21*** (0.06)	0.00 (0.00)
closeness centrality unweighted:nontrivial						0.06 (0.04)	
closeness centrality weighted					0.22*** (0.05)		0.39*** (0.06)
clustering coefficient				-0.04 (0.06)		-0.03 (0.06)	-0.18*** (0.06)
days since first post			0.06 (0.06)	0.01 (0.05)	0.01 (0.05)	0.01 (0.05)	0.09 (0.06)
degree incoming			0.04 (0.06)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
degree outgoing				0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.07 (0.11)
degree total				0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.05 (0.10)
num posts			-0.03 (0.06)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	-0.17 (0.10)
num subjects				0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	-0.05 (0.06)
pagerank weighted					0.00 (0.00)		-0.17 (0.19)
satoshi distance							0.00 (0.00)
satoshi distance inf							-0.04 (0.05)
satoshi pagerank weighted							0.18 (0.19)
R2	0.00	0.02	0.04	0.10	0.10	0.11	0.16
ElasticNet CV MSE:	1.01	0.99	0.99	0.94	0.95	0.94	0.94
BIC	1072	1069	1082	1078	1082	1080	1084
N	376	376	376	376	376	376	376
Adjusted-R2	0.00	0.02	0.03	0.09	0.08	0.09	0.13
Condition Number	1.00	1.00	2.05	nan	147295588.13	nan	nan

Table 2: Log(Severity) model OLS parameter estimates with heteroskedasticity robust SE after ElasticNet variable selection

	Activity	Nontrivial	Satoshi	Network	Weighted	Network*Nontrivial	All
Intercept	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
nontrivial					-0.03		
					(0.05)		
closeness centrality unweighted				0.00		0.00	0.15**
				(0.00)		(0.00)	(0.06)
closeness centrality unweighted:nontrivial						0.00	
						(0.00)	
closeness centrality weighted					0.30***		0.00
					(0.05)		(0.00)
clustering coefficient				0.00		0.00	0.23***
				(0.00)		(0.00)	(0.05)
days since first post					-0.02		-0.01
					(0.06)		(0.06)
degree incoming				0.00	-0.09*	0.00	-0.09
				(0.00)	(0.05)	(0.00)	(0.07)
degree outgoing					0.09		0.00
					(0.09)		(0.00)
degree total					-0.04		0.00
					(0.03)		(0.00)
num posts					-0.16		-0.08
					(0.10)		(0.07)
num subjects					-0.00		-0.01
					(0.06)		(0.06)
pagerank weighted					0.36***		0.55***
					(0.07)		(0.18)
satoshi distance							-0.11
							(0.11)
satoshi distance inf							0.08
							(0.10)
satoshi pagerank weighted							-0.22
							(0.17)
R2	0.00	0.00	0.00	0.00	0.22	0.00	0.27
ElasticNet CV MSE:	1.01	1.01	1.01	0.94	0.91	0.94	0.90
BIC	1072	1072	1072	1090	1028	1096	1024
N	376	376	376	376	376	376	376
Adjusted-R2	0.00	0.00	0.00	-0.01	0.20	-0.01	0.25
Condition Number	1.00	1.00	1.00	1.80	147295588.13	2.17	nan