Characterizing Speed and Scale of Cryptocurrency Discussion Spread on Reddit

Maria Glenski
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana
mglenski@nd.edu

Emily Saldanha
Data Sciences and Analytics Group
Pacific Northwest National Laboratory
Richland, Washington
emily.saldanha@pnnl.gov

Svitlana Volkova
Data Sciences and Analytics Group
Pacific Northwest National Laboratory
Richland, Washington
svitlana.volkova@pnnl.gov

ABSTRACT

Cryptocurrencies are a novel and disruptive technology that has prompted a new approach to how currencies work in the modern economy. As such, online discussions related to cryptocurrencies often go beyond posts about the technology and underlying architecture of the various coins, to subjective speculations of price fluctuations and predictions. Furthermore, online discussions, potentially driven by foreign adversaries, criminals or hackers, can have a significant impact on our economy and national security if spread at scale.

This paper is the first to qualitatively measure and contrast discussion growth about three popular cryptocurrencies with key distinctions in motivation, usage, and implementation – Bitcoin, Ethereum, and Monero on Reddit. More specifically, we measure how discussions relevant to these coins spread in online social environments – how deep and how wide they go, how long they last, how many people they reach, etc. More importantly, we compare user behavior patterns between the focused community of the official coin subreddits and the general community across Reddit as a whole. Our Reddit sample covers three years of data between 2015 and 2018 and includes a time period of a record high Bitcoin price rise.[1]

Our results demonstrate that while the largest discussions on Reddit are focused on Bitcoin, posts about Monero (a cryptocurrency often used by criminals for illegal transactions on the Dark Web[2]) start discussions that are typically longer and wider. Bitcoin posts trigger subsequent discussion more immediately but Monero posts are more likely to trigger a longer lasting discussion. We find that moderately subjective posts across all three coins trigger larger, longer, and more viral discussion cascades within both focused and general communities on Reddit. Our analysis aims to bring the awareness to online discussion spread relevant to cryptocurrencies in addition to informing models for forecasting cryptocurrency price that rely on discussions in social media.

[1]http://fortune.com/2017/12/17/bitcoin-record-high-short-of-20000/
[2]https://www.deepdotweb.com/2017/09/20/dhs-says-darknet-criminals-switching-bitcoin-monero/

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313702

CCS CONCEPTS

• **Networks** *Online social networks*.

KEYWORDS

discussion spread; information cascades; social networks; Reddit

ACM Reference Format:
Maria Glenski, Emily Saldanha, and Svitlana Volkova. 2019. Characterizing Speed and Scale of Cryptocurrency Discussion Spread on Reddit. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3308558.3313702

1 INTRODUCTION

Cryptocurrencies are an emerging and disruptive technology. The coins themselves, the underlying software, and the surrounding environment of social and trading behavior are highly volatile and evolve quickly. Cryptocurrencies are disproportionately used by criminals and hackers, and their use has political and economic implications. The way information about these novel distributive technologies spreads across online social platforms shapes online and offline discussions, and could potentially influence peoples' beliefs and decision making. Thus, it is crucial to understand, explain, and anticipate the social behavior and communication patterns in the social environments surrounding cryptocurrencies to understand this phenomena and devise appropriate responses [22, 30, 35].

A preliminary analysis of potential cryptocurrencies of interest identified Bitcoin (BTC), Ethereum (ETH), and Monero (XMR) as the top three cryptocurrencies in terms of both developer interest on GitHub and Bitbucket and community interest on social media platforms – Twitter, Reddit and Facebook as illustrated in Figure 1. For that analysis, we collected data from CoinGecko for 1,742 cryptocurrencies.[3] and compared coins in terms of developer interest features (a measure of activity in public repos on GitHub and Bitbucket) and community interest features (a measure of discussions and popularity on social media platforms) released by CoinGecko. *Bitcoin* is a market-leading cryptocurrency, which laid the foundation for blockchain's transparent digital ledger system (blockchain). *Ethereum* is a cryptocurrency as well as a blockchain development platform, where developers may build their own automated smart contracts. *Monero* is a cryptocurrency similar to Bitcoin but with added security and anonymity features which enable more privacy in transactions. Monero is often used by criminals for illegal transactions on the Dark Web[2].

[3]https://www.coingecko.com/en/coins/all

Figure 1: Cryptocurrencies plotted by developer and community interest according to CoinGecko.

To motivate our analysis we present discussion growth on Reddit between January 2015 and January 2018, covering a period that includes the record high Bitcoin price rise at the end of 2017, in the official r/bitcoin subreddit aligned with related real-world events in Figure 2. We observe a significant increase in Reddit discussion volume during the period of the record Bitcoin price increase. During that time, online discussions could be influenced by adversaries [24] who could potentially manipulate people's opinions [3, 10] regarding buying or selling Bitcoin. The same is true for other cryptocurrencies, for example, we discover high correlations measured as Normalized Mutual Information (NMI) between Reddit discussion volume and coin price for Bitcoin (NMI = 0.64), Ethereum (NMI = 0.89) and Monero (NMI = 0.82).

Our main contribution is a measurement-driven analysis of cryptocurrency discussion spread for three popular coins within two Reddit sub-populations: the *Official Subreddit* communities and a broader *Crypto-Ecosystem* comprised of multiple crypto-related communities. We measure the *speed* of cryptocurrency discussion spread to understand how quickly discussions begin (initial delay or responsiveness), how they grow over time, and how quickly they end (discussion lifetime). We evaluate the *scale* of cryptocurrency discussion spread by measuring discussion volume, audience size, and structural properties of discussion threads *e.g.,* the max-breadth, max-depth, and structural virality [16].

2 RELATED WORK

Information cascades in online social networks (OSNs) are a phenomena in which the idea becomes widely adopted due to influence of social network neighbors and has been widely studied [4, 5, 17, 18, 27, 28, 33]. Researchers presented analyses on how different types of information spread in different social environments e.g., how images spread on Flickr [6, 7], pins (images) on Pinterest [19, 20], videos on YouTube [36], reviews on Yelp [21], URLs on Yahoo [17], posts on Weibo [34], hashtags and URLs on Twitter [26, 31], posts and memes on Facebook [1, 9, 11] etc. Another line of work focused on predicting cascade properties e.g., size, depth [2, 8, 23, 25]

Unlike prior work that focused on how a specific piece of content e.g., an image, URL etc. spreads in a single social environment and what factors drive such information spread, we focus on analyzing threaded discussion cascades relevant to the cryptocurrency domain on Reddit. Our analysis is motivated by the phenomena where online discussion spread could potentially influence (and be influenced) by real-world external events e.g., market crashes and coin prices. The most similar work to ours is [12]. The authors focused on the most popular subreddits and performed the analysis of conversation patterns on Reddit by only considering one measurement of speed (responsiveness), and two scale measurements (volume and virality). We go beyond that: we empirically evaluate and contrast speed and scale of discussion spread related to cryptocurrencies using multiple measurements within the *Official Subreddit* communities and discussions within a broader *Crypto-Ecosystem.*

3 REDDIT DATA

In this study we chose to focus on analyzing discussion spread on Reddit for several reasons. Reddit is a highly popular social news aggregator[4] that allows communities of users to share and discuss information, opinions, and entertainment media. There are 330 million users active in 140 thousand communities on Reddit who post 2.8. million comments daily, and perform 11 million posts and 11 billion page views monthly[5]. Moreover, reconstructing cascades for Twitter and Facebook requires either access to the full data (not just a sample [13]), the collection of all followers to be able to reconstruct the cascades, or the application of other methods for correcting information cascades for missing data [32]. Unlike these social networks, Reddit post and comment data is public[6] and contains full discussion cascades.

Cryptocurrency Official Subreddits. The *Official Subreddit* dataset comprises all discussions posted to r/bitcoin,r/ethereum, and r/monero subreddits over a period of three years from 2015 through 2017. These three subreddits are the official communities for each of our cryptocurrencies of interest and had substantial traffic of, on average, 3.6K, 500, and 380 comments posted each day, for Bitcoin, Ethereum and Monero, respectively. As these communities are explicitly tied to each of the cryptocurrencies, we label each discussion with the subreddit's respective coin. This resulted in a total of 212,302 Bitcoin, 41,792 Ethereum, and 15,035 Monero discussions.

Cryptocurrency Ecosystems. To collect relevant discussions across the Reddit community in general for the *Crypto-Ecosystem* data, we collected the full comment threads for posts that were (a) submitted from January 2015 through August 2017 and (b) contained either keywords related to Bitcoin ('#btc', '#BTC', '#bitcoin', '@bitcoin', 'bitcoin'), Ethereum ('#eth', '#ETH', '#ethereum', '@ethereum', 'ethereum'), or Monero ('#xmr', '#XMR', '#monero', '@monero', 'monero') in the title or text (if the post includes text). This resulted in 38,118 posts and comments for Bitcoin, 8,989 for Ethereum, and 4,381 for Monero that were submitted from January 2015 through August 2017. These discussions are not partitioned by subreddit and, intuitively, discussions can be started by posts that include keywords for multiple coins. To avoid bias from one coin affecting analysis of another, we restrict the dataset to include only those discussions where the initial post included keywords for a single coin.

[4]The 18th most popular site globally and 5th most popular within the U.S. according to Alexa.com: https://www.alexa.com/topsites
[5]https://expandedramblings.com/index.php/reddit-stats/
[6]An archive of Reddit posts and comments is publicly available at https://files.pushshift.io/reddit/ and covers the time period of 2005 through 2018 as of February 2019.

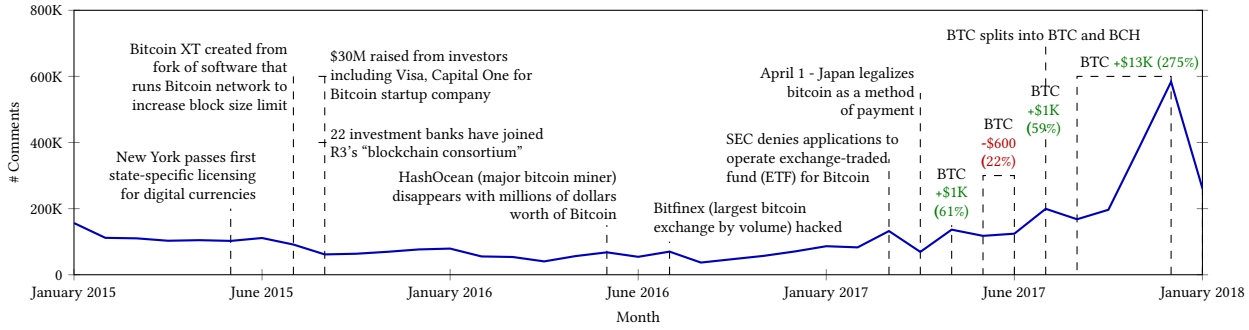


Figure 2: Motivation: discussion growth on Reddit for the r/bitcoin official subreddit related to real-world events and a record high Bitcoin price rise at the end of 2017.

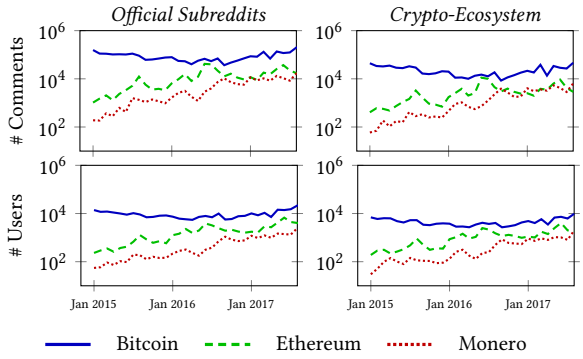


Figure 3: Discussion activity over time as the number of comments (above) and users (below) for each cryptocurrency of interest. Y-axes are presented on a log scale.

This resulted in 35,282 discussions for Bitcoin, 6,949 for Ethereum, and 3,324 for Monero. This restriction did not significantly reduce the sample size overall or for any one coin.

Figure 3 illustrates the volumes of comments posted (above) and users active (below) in cryptocurrency-related discussions each month within two datasets described above. To replicate our results, one could collect comments from the Reddit archive publicly available at https://files.pushshift.io/reddit/ and extract all comments posted to the official subreddits or with related keywords present in the body of the comment, as described above. Interestingly, we see that the volumes of active users and comments posted to the official subreddits, where we did not restrict posts to only those that included the set of keywords, follow very similar trends to the corresponding sets of discussions within the Crypto-Ecosystem dataset.

4 METHODOLOGY

On Reddit, users may contribute to discussions in several ways:

- starting a discussion by authoring a post which must be submitted to a specific community (subreddit),
- posting a comment in reply to a post,
- posting a comment in reply to another comment,
- voting (positively or negatively) on posts or comments.

These discussions can be viewed as tree structures or information cascades (aka comment trees on Reddit) [4, 12, 18] wherein the root node represents the initial post, and other nodes in the tree represent comments to the initial post. We define a discussion cascade as an undirected tree $T = (V, E)$, where V represent all messages (the original post and the follow-up comments in the discussion thread), and E represent the set of edges that connect messages linked by in-reply (or commenting) actions. Figure 4 presents the discussion cascade for an example post and ten comments. We focus on measuring discussion growth using comment trees within focused communities of the official subreddits and across multiple communities in the wider cryptocurrency ecosystems and characterize the speed, scale, and a high-level content analysis of cryptocurrency related discussions on Reddit for three coins of interest – Bitcoin, Ethereum, and Monero. Before studying discussion spread, similar to [5, 27], we discuss the affect of the topological structure of the network, and characterize the population of contributors as outlined below.

First, we examine the *population of contributors* who participated in these discussions in terms of how active they are, *e.g., are we analyzing the behavior of a large population with varied frequencies*

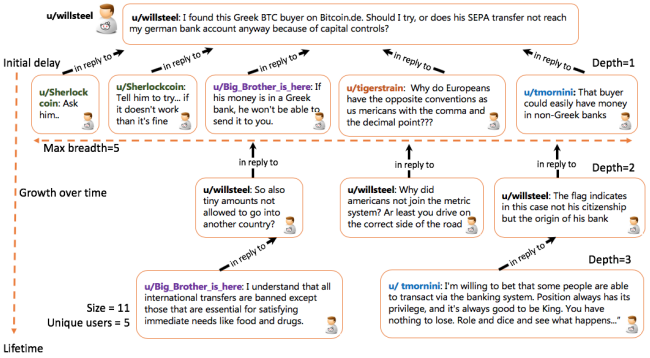


Figure 4: Discussion cascade (aka comment tree on Reddit) illustrated for a post with 10 comments. Discussion speed measurements – initial delay, growth over time, lifetime and discussion scale measurements – volume, audience and structural properties of information cascades are shown.

of activity or a primarily homogenous, densely connected network of highly active users? For that, we look at the mean averages and respective 95% confidence intervals for the volumes of discussions related to each coin that users initiate, *i.e.,* by submitting the initial post, or participating in the discussion through posts and comments.

Next, we examine *topology of user networks* (the forest of cascade trees) constructed created from the reply-links in the discussion threads as shown in Figure 4. We define a reply-link as follows: reply-link <*A*,*B*> exists if user *A* posted a comment in reply to a comment or initial post authored by user *B*. The goal is to determine how closely connected communities contributing to cryptocurrency discussions are, *e.g., are users discussing cryptocurrencies back and forth with the same set of users or engage new users?* The clustering coefficients of the resulting user networks measure the extent to which nodes cluster together in communities within each network.

Finally, we consider the *prevalence of posts that do not trigger any additional commentary* for each coin. We call these posts *ineffective posts*. Conversely, we define *effective posts* as posts that prompt at least one subsequent comment. We compare the relative rate of ineffective posts for each cryptocurrency of interest across both datasets. We examine whether there are consistent patterns in effectiveness of posts related to a given coin or patterns being pairs or sets of coins across datasets. To mitigate bias on measurements of growth and speed analyses from these ineffective posts, we remove them from subsequent analyses in all but one case: when we consider the size of discussions.

4.1 Measuring Speed of Discussion Spread

There are several ways to characterize the speed of discussion growth. In this work we choose to focus on the speed of the three main temporal phases of discussions to characterize discussions related to each coin based on:

(1) how quickly they begin,
(2) the speed with which they grow over time,
(3) how quickly they end.

Initial Delays (aka Responsiveness). The first speed-characteristic we consider the delay between when the initial post is submitted and when the first comment in the discussion is posted in response. This relies on a minimal discussion threshold of two contributions (*i.e.,* one post and one comment). We call this the *initial delay* of the discussion which we operationalize as a measure of the discussion's initial growth rate. As such, we only include discussions that grew beyond the initial post. To compare the average behavior in terms of initial growth, we compare the mean and median values for each coin of interest and determine whether there are statistically significant differences with Mann Whitney U (MWU) tests [29]. Finally, we plot the complementary cumulative distribution functions (CCDFs) for each coin. Complementary cumulative distribution functions plot the percentages of cascades with delays at least as long as a given delay. Through this analysis we characterize how quickly a post triggers a discussion thread.

Growth over time. Next, we consider *the time to reach each depth* in the discussion thread. That is, the delay between submission of the initial, root post and the first comment posted within each depth in the discussion tree. For this measure, we plot the mean average delay to reach each depth with associated 95% confidence intervals for each of the coins. As the mean tends to be influenced by extreme outliers, we also examine the median delay to reach each depth and determine whether there are more robust trends or whether trends remain consistent across both mean and median delays by depth. Similarly, we also compare the mean and median *size by depth*. That is, how quickly (as determined by the depth within the discussion) do the discussions grow at each stage. Through this analysis we identify whether discussions grow consistently or sporadically.

Lifetimes. Thirdly, we consider the *lifetime* which we define as the length of time a discussion is active, *i.e.,* the delay between the submission of the root post and the last comment in the discussion, to characterize how quickly discussions end. Similarly to the analysis we perform related to the initial delays, we compare and contrast the mean and median values and plot the CCDFs for each coin across datasets. Again, we test for significant differences in discussion lifetimes between coins with MWU tests as our approach to partition discussion cascades by coin results in independent sets of discussions.

4.2 Measuring Scale of Discussion Spread

Next, we measure the scale of cryptocurrency discussion growth. We focus on summarizing discussions in terms of the discussion volume, audience size, and structural properties e.g., max breath, max depth and structural virality.

Volume. The first measure of discussion scale we consider is the final size of the discussion, *i.e.,* the final volume of nodes within the discussion tree. Therefore, we define the size as the number of comments plus the initial post that prompted the discussion. As such, *ineffective posts* noted previously are discussions of size 1. We present CCDFs for discussion sizes for each coin and contrast behavior across two datasets. We also present and compare the mean, median, and maximum discussion sizes. MWU tests once again allow us to determine significant differences between coins. Intuitively, we would expect discussions posted to larger subreddits to have a higher maximum size and larger range, in general, for discussions sizes. This is because with larger subreddit sizes (and, thus, more contributors), there are more potential users who may be inclined to contribute to each discussion. However, while in the larger subreddits, there are more potential participants that could drive a discussion to a much larger size, there is also a higher influx of content to the larger subreddits where newer content quickly pushes old discussions down and out of view. This flow of new content may motivate users to join a more recently started or active conversation before they even see the older conversations.

Audience size. We can also consider size in terms of the audience or contributors to the discussion rather than the contributions (posts and comments). We define the population as the number of contributors (users) who participated in a discussion thread by submitting the initial post or a subsequent comment within the thread. We can then summarize the overall population sizes and compare the mean and median values across coins and datasets. Once again, because of the independence of the samples for each coin, we can identify whether significant differences in population sizes exist with MWU tests comparing each of the coins to another.

We also compare the maximum values to consider the upper range of populations who discuss each coin. This analysis provides an overview of the static property, population overall. We also consider the populations within the discussions' cascades and at varying phases of the discussion. To do so, we present the mean and median numbers of users who contribute within a given depth and focus on the initial depths within the discussion: *is there a consistent population contributing regularly or a large population whose contributions are concentrated within a single depth?*

Structural properties. Finally, we consider scale-based characteristics in terms of the structure of the discussion cascades. Here, we compare several static properties of discussion structure:

(1) the *structural virality* or Weiner Index (WI) of the cascade, a measurement of how closely nodes are connected within the discussion [16];
(2) the *max-depth*, *i.e.,* the length of the longest path from the root (initial post in the discussion) to a leaf node in the cascade;
(3) the *max-breadth*, *i.e.,* the largest number of comments within a single depth of the discussion tree.

Again, we compare the CCDF plots across coins and analyze first order statistics for each of the structural properties, with significant differences identified with MWU tests. Max-Depth and Max-Breadth provide insight into the primary dimensions of the cascade individually to identify whether discussions are typically wide or long. Structural virality, on the other hand, provides a more easily interpretable measure of the structure as a whole. Weiner index become minimum when all comments are directly added to the root, and maximum when the tree becomes a chain. In other words, *do discussions follow a "viral" structure of multiple branches off of branches within the discussion or a structure that is more reminiscent of a "broadcast" that is wide but not very deep?* To capture a dynamic measure of the structure as the discussion grew rather than the static properties represented with Max-Depth, Max-Breadth, and structural virality, we also plot the mean average and median breadth by depth. Here, we see how the breadth or width of the discussion cascade grows and shrinks over the initial depths. *Are these patterns of growth monotonic (*i.e., *consistently increasing or consistently decreasing), or do they oscillate between growing and shrinking at increasing depths?*

4.3 Content Subjectivity and Discussion Spread

We focus on the subjectivity of the initial posts and their effect on the discussion spread. To get post subjectivity we rely on TextBlob toolkit[7]. The subjectivity score is a float between 0 and 1, where 0 is very objective and 1 is very subjective. First, we compare the subjectivity of successful versus unsuccessful discussions at a very high level dichotomy: *did the initial post trigger subsequent discussion or did the discussion consist solely of the initial post?* We plot the kernel density estimations (KDE) of what we call *effective posts* that initiated discussions of at least one additional comment and *ineffective posts* that failed to initiate a discussion for each of the coins. We can then compare these KDE plots to identify changes in distributions of subjectivity – *do ineffective and effective posts have significantly different distributions of subjectivity?* Then, to gather a more nuanced view of how the initial post in discussions influence discussion growth, we compare correlations measured using normalized mutual information for each of the static discussion properties (virality, max-depth, max-breadth, size, and population) with the subjectivity of the initial (or "root") post in the discussion.

[7] https://textblob.readthedocs.io/en/dev/

Table 1: Average number of cascades a user initiated, *i.e.,* as the author of the initial post, or participated in as the author of an initial post or subsequent comment.

	Official Subreddits		*Crypto-Ecosystem*	
Coin	Initiated	Participated	Initiated	Participated
Bitcoin	3.79 ± 1.20	11.27 ± 1.19	1.90 ± 0.19	6.74 ± 0.54
Ethereum	3.23 ± 0.96	7.56 ± 1.13	1.80 ± 0.09	4.24 ± 0.42
Monero	2.90 ± 0.75	8.50 ± 1.22	1.84 ± 0.17	5.75 ± 0.76

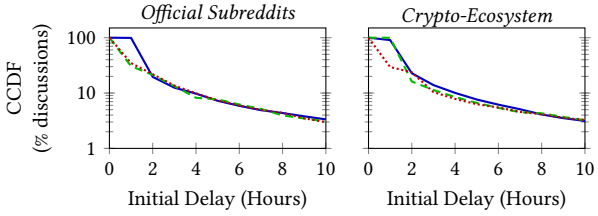
5 RESULTS

Before we measure the speed and scale of information spread in cryptocurrency discussion growth, we first examine several influential characteristics of each dataset and the potential impacts on our analysis. Here, we examine the population of contributors who participated in these discussions, the user-networks for each dataset, and the prevalence of ineffective posts – posts that did not spread.

First, we look at the mean averages and respective 95% confidence intervals for the volumes of discussions related to each coin that users initiate, *i.e.,* by submitting the initial post, or participate in (through posts and/or comments), shown in Table 1. We find that *users are more active within a single subreddit dedicated to* i.e., *the Official Subreddits domain than within discussions focused on the same coin that are started in multiple related subreddits.* Thus, there should be less bias from a single user or group of users on the results of discussion growth in the Crypto-Ecosystem dataset. Intuitively, this composition of contributing users makes sense as users typically browse, subscribe, or otherwise follow specific subreddits based on the topics they cover than individual posts that discuss a topic of interest. Further, a recent study found that users tend to focus their interactions in a small number of subreddits [14].

Next, we consider the networks created with the reply-links to examine how closely connected these communities are. *Are users discussing back and forth with the same set of users or are most interactions with new users?* We find that users interact with, on average, 13 users in the r/Bitcoin subreddit, 8 users in the r/Ethereum and r/Monero subreddits within the Official Subreddits domain. Within the Crypto-Ecosystem domain, users who participate in Bitcoin-related discussions interact with approximately 9 users and we see a similar drop as we do in the Official Subreddits to around 5 and 7 for Ethereum and Monero, respectively. The clustering coefficients are also fairly low for both the Official Subreddits – 0.13 and Crypto-Ecosystem – 0.12 datasets. As shown in Table 1, users, on average, participate in a fairly low number of discussions (between 8 and 11 for Official Subreddits and between 4 and 7 for Crypto-Ecosystem) overall. Users are not densely connected through frequency in discussions or replies to specific users.

Next, we look at the prevalence of posts that do *not* trigger any additional commentary for each coin across datasets. We call these



	Official Subreddits			Crypto-Ecosystem		
coin	mean	median	max	mean	median	max
Bitcoin	4.61	11.17	6.31	5.26	15.25	6.00
Ethereum	7.60	27.60	6.28	8.12	36.37	5.00
Monero	2.85	20.98	4.38	2.56	23.36	3.00

Figure 5: CCDFs for the initial delays of discussions (above) and the mean delay in hours, the median delay in minutes, and the max delay in months (below).

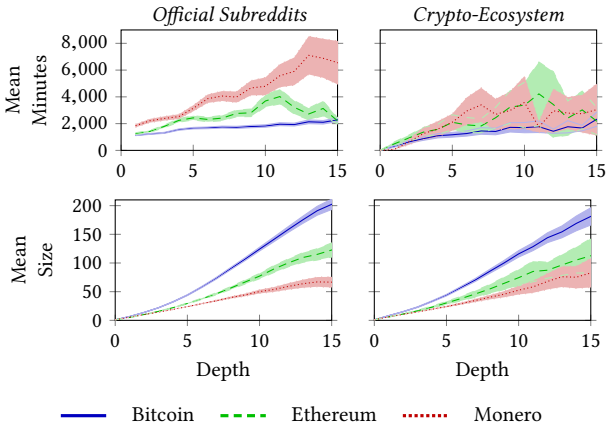


Figure 6: Discussion growth over time measured as the mean average minutes to reach a given depth (above) and the mean size of the discussion when it reaches a given depth (below). 95% confidence intervals are represented with shaded bands.

ineffective posts. We remove these posts from subsequent analyses in all but one case: when we consider the size of discussions. We find an interesting trend across both datasets where Bitcoin and Ethereum have similar rates of ineffective posts (31% and 33%, respectively, for the Official Subreddits and 8% and 9% for the Crypto-Ecosystem) and Monero has much lower rates (13% and 2%). The relative presence of ineffective posts related to the more popular Bitcoin and Ethereum coins are twice to five times as high as found in Monero discussions.

5.1 Speed of Discussion Spread

Here we examine the speed of discussion growth across the three coins of interest and two datasets. We aim to answer three main research questions: *how quickly do discussions begin, grow over time and end*. By answering these three questions we provide a well-rounded overview of both dynamic and static properties of the speed of discussion growth. As we noted previously, we removed what we call *ineffective posts*, posts that failed to trigger a discussion, from the following analysis.

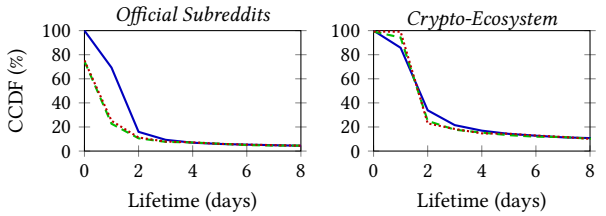
5.1.1 Initial Delays. First, we examine how quickly discussions begin in Figure 5. As a result of the size of the r/bitcoin community and the number of new posts that are submitted to the subreddit, the outliers (e.g. extremely long reaction times from posts that either slowly made their way from the "new queue" to the main subreddit page or were possibly found through the search functionality long after the initial submission) may highly influence mean average delays. On the other hand, the median delay is less influenced by outliers. Bitcoin discussions in the official subreddit have the shortest initial delay of approximately 11 minutes. That is, the median delay between a post submitted to r/bitcoin and that post's first comment is around 11 minutes. In comparison it takes at least 20 minutes for a typical post to Monero to attract its first comment and over 27 minutes for Ethereum. We see that the same ordering is found in the Crypto-Ecosystem. Although there are larger initial delays, we still find that Bitcoin discussions begin the fastest, followed by Monero. Ethereum discussions are the slowest to start. But does this trend follow when we consider growth beyond the first comment?

5.1.2 Growth over Time. Now, we expand beyond the initial delay to the average time it takes for discussions to grow, in general. We examine the speed of discussion growth through the average time to reach each depth, the average size of cascade as each depth is reached and the population size within discussions as they grow.

First, we plot the mean minutes elapsed before discussions reached a given depth in Figure 6 (above) with 95% confidence intervals represented with shaded bands. Although the scale decreases greatly when medians are used instead of mean averages, we find that the trends between the three coins remain consistent. Bitcoin discussions grow the fastest followed by Ethereum then Monero. Although we see some overlap in the 95% confidence intervals when we plot the mean minutes to reach each depth for the Crypto-Ecosystem, the ordering remains consistent.

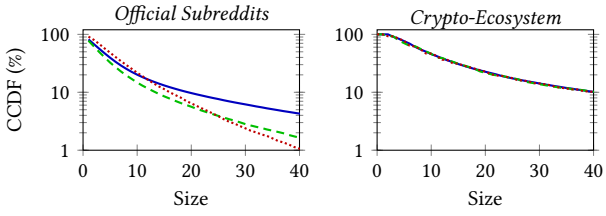
If we consider the size of discussions by the time each depth in the tree is reached, as illustrated in Figure 6 (below), we see that Bitcoin discussions are the largest at each depth and more efficiently increase the number of comments at each depth than Ethereum and Monero across both the datasets. Although the order is flipped (BTC > ETH > XMR), this finding is consistent with the pattern found in average time to reach depth in that Bitcoin shows the fastest growth to largest sizes. This pattern is consistent not only across both datasets but in both median and mean discussion size at each stage of evolution, as determined by the depth in the discussion tree.

5.1.3 Discussion Lifetimes. Finally, we look at how long discussions are active. The median and maximum periods of activity in discussions, *i.e.*, lifetimes, are presented in the Figure 7. We find that Monero discussions have the largest median lifetimes but not the largest maximum lifetimes. Discussions with the largest possible lifetimes are related to the Ethereum cryptocurrency. Interestingly, Bitcoin discussions have the lowest median lifetimes of all three



	Official Subreddits			*Crypto-Ecosystem*		
coin	mean	median	max	mean	median	max
Bitcoin	107.27	6.69	6.43	161.62	13.19	6.00
Ethereum	140.88	11.85	11.12	182.50	17.99	10.00
Monero	182.74	19.18	6.43	178.27	24.11	6.00

Figure 7: CCDFs for the lifetime of discussions (above), the mean, median (in hours), the max lifetimes in months (below).



	Official Subreddits			*Crypto-Ecosystem*		
coin	mean	median	max	mean	median	max
Bitcoin	18.02	6	7,867	20.63	8	2,843
Ethereum	12.74	5	3,816	15.66	7	2,111
Monero	26.98	16	942	17.55	11	692

Figure 8: CCDFs for discussion sizes (above) with the y-axis on a log scale, and the mean, median, and max sizes (below).

coins. In Figure 7, we illustrate the CCDF plots for discussion lifetimes focused on a window of up to eight days which captures 90-95% of all discussions. We see that there is a clear distinction between the complementary cumulative distribution functions for Bitcoin and the other two coins in the Official Subreddits dataset.

Bitcoin has the largest subreddit in terms of content submitted and users who subscribe or contribute, and, as shown in Figure 3 a larger and consistent stream of discussions and comments in both the Official Subreddits and Crypto-Ecosystem datasets. With the consistent flow of new content, users may be more inclined to contribute to a newer discussion than join and continue an older discussion that may no longer be visible because of its score or rank pushing it below a threshold of highly visible ranks. As rank bias influences which content gets voted on and seen by new users, this will also influence which discussions users decide to engage with.

As we noted previously, Monero is the least popular or well-known of the three coins. This is reflected in the smallest subreddit (both in terms of subscribers, contributors, and content) and the smallest sample of posts within the Crypto-Ecosystem. The lower amount and relative frequency, as opposed to the more popular Bitcoin and Ethereum cryptocurrencies, of new content submitted about Monero that would disrupt an existing discussion has probably influenced the larger median lifetimes. With a smaller core community, users who comment in the Monero subreddit may also be more committed to continuing conversations with other contributors, as opposed to users who are tangentially interested in who may visit the larger bitcoin subreddit and make one-off comments on posts they browse. Similarly, the posts related to Monero in the Crypto-Ecosystem dataset may be submitted to smaller subreddits or appeal to a smaller population.

5.2 Scale of Discussion Spread

In this section we present the results on the scale of cryptocurrency discussions, in particular, *how large* discussions grow and the structure of discussions overall and as they develop. In this section, we include the discussions comprised of the *ineffective posts* that we removed from the previous analysis on speed.

Table 2: Discussion audience size measured as median and max volumes of users who participated in each thread.

	Official Subreddits			*Crypto-Ecosystem*		
coin	mean	median	max	mean	median	max
BTC	7.30	3	3,037	11.00	5	685
ETH	5.05	2	707	7.63	4	707
XMR	7.09	5	177	8.63	6	370

5.2.1 Discussion Size. To measure discussion scale we consider the final size of the discussion, *i.e.,* the final volume of nodes within the discussion tree, and plot the CCDFs for discussion sizes for each coin in Figure 8. Across both datasets, the window of sizes up to 40 captures 90 to 99% of all cascades. We see that sizes are more heavily concentrated in lower range of sizes for the Official Subreddits dataset. The CCDF for the r/Bitcoin subreddit rises above the others, indicating more discussions of a larger final size in r/Bitcoin than in r/Ethereum and r/Monero.

When we compare the mean, median, and maximum discussion sizes in the table (below) in Figure 8, we see that the range of r/Bitcoin is much more extended. In contrast, the median size of a Bitcoin cascade is smaller than that of a Monero cascade. Further, Mann Whitney U comparisons of the distributions of each coin's discussion sizes find that, on average, Ethereum discussions are smaller than Bitcoin and Monero discussions ($p < 0.01$), and Bitcoin discussions are smaller than Monero discussions ($p < 0.01$). Although some Bitcoin discussions are much larger than Monero discussions, most are not. Again, this may be due to the size of the community. In the larger subreddits, there are more potential participants that could drive a discussion to a much larger size but there is also a higher influx of content to the larger subreddits where newer content quickly pushes old discussions down and out of view.

5.2.2 Discussion Audience. We can also consider size in terms of the audience or contributors to the discussion rather than the contributions (posts and comments). The overall population sizes are

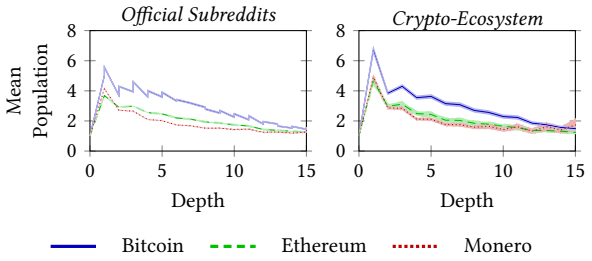


Figure 9: Discussion audience within depth measured as the mean population size at each depth with the 95% confidence interval represented with shaded bands (above), and the median contributor count at each depth (below).

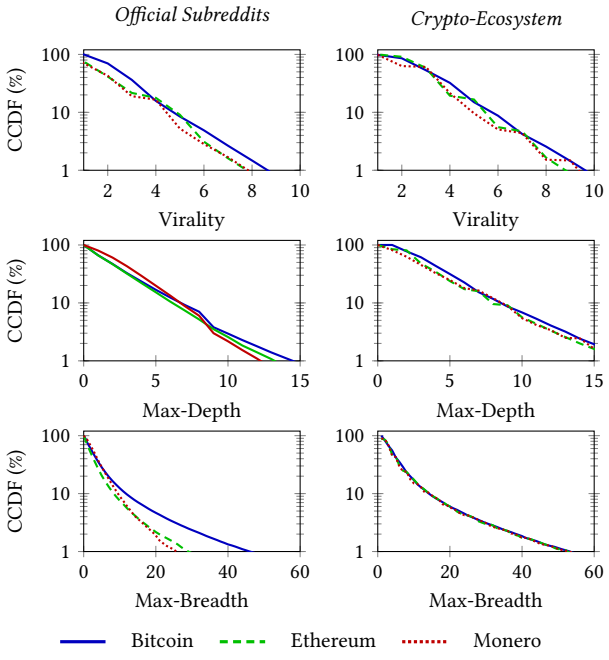
summarized in Table 2. Again, we present the median and maximum overall population (or audience) sizes for discussions related to each coin. The typical discussion is held between 2 and 6 users.

Among the official subreddits, the population sizes are quite similar but in the Crypto-Ecosystem when we consider only those discussions that explicitly reference the coin in the initial post, we see that Bitcoin and Ethereum have a median of five and four users participating, respectively, while Monero has a median of seven users. On the other hand, Monero has the lowest maximum population size while Bitcoin has the largest. The ordering (ETH < BTC < XMR) in terms of typical numbers of participants is also found when we compare the distributions of population sizes with Mann Whitney U tests ($p < 0.01$). As with size of contributions, Monero typically has larger populations in discussions but not the largest possible.

In Figure 9, we observe that the mean population size at each depth differs significantly between coins within the first 15 depths. Across both datasets, Bitcoin discussions have more users participating at each depth. In the Official Subreddits, discussions in r/Ethereum have more users at depths greater than two. That is, discussions in r/Monero on average have slightly more users who respond directly to the initial post than those in r/Ethereum, however this trend is flipped for comments in response to other comments. In the Crypto-Ecosystem dataset, we see an overlap between Ethereum and Monero discussions. When we consider the median population within each depth, we see more similar behavior among the three coins, with spikes above average at several depths for the Bitcoin discussions.

5.2.3 Discussion Structure. Finally, we look at how the scale of the discussion and evolution of the discussion in terms of the structure of the discussion tree. We compare static properties of discussions overall with CCDF plots of virality, max-depth, and max-breadth in Figure 10. As we see in the center row, few discussions have any one given chain grow beyond 5 comments. Less than 1% of discussions have a chain that is at least 15 comments long in the official subreddits, approximately 1-2% has such a chain in the Crypto-Ecosystem dataset.

We highlight the median values for each of the structural properties in Figure 10. We see that discussions tend to have fewer depths with a moderate (at most 3 to 6) number of comments at the most populated depth. Mann Whitney U comparisons of the distributions



	Official Subreddits			*Crypto-Ecosystem*		
	BTC	ETH	XMR	BTC	ETH	XMR
Virality	1.6	1.5	2.3	2.3	2.2	2.8
Max-Depth	1	1	3	3	3	4
Max-Breadth	3	3	6	4	3	4

Figure 10: CCDFs for structural measurements of discussion cascades with a log-scale used for each y-axis (above) and the median values of each property (below).

find that across both datasets, Ethereum has the shortest, narrowest, and least viral of all the coins ($p < 0.01$). That is, these discussions appear more like a broadcast of information than a viral cascade – the initial post is typically followed by a single layer of comments who respond directly to that initial post. Monero discussions have the most "viral" of all discussions, on average, with larger maximum depths and breadths, on average (MWU $p < 0.01$).

When we consider the average comment count (breadth) at each depth however, we see that Bitcoin discussions exhibit larger breadths at each of the initial depths. Figure 11 highlights the mean (above) and median (below) breadths found at each of the initial 15 depths. This illustrates the key differences between coins when we consider discussions as they grow versus the overall discussion after completion. While the maximum breadth of cascades is typically larger in Monero discussions, we see more consistently large breadths in Bitcoin. As we saw in the CCDF plots in Figure 10, Ethereum and Monero discussions exhibit similar patterns of behavior.

5.3 Post Subjectivity and Discussion Spread

In this section, we go beyond the static and dynamic measures of growth and discussion behavior to focus on the *contextual* features of the initial posts that either triggers or fails to trigger a discussion

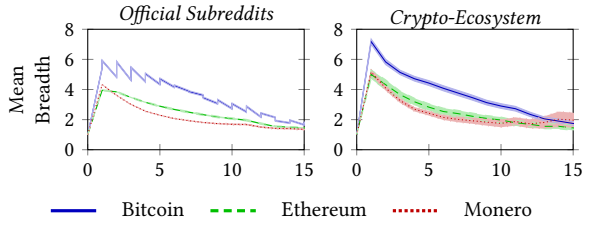


Figure 11: The mean breadth by depth with the 95% confidence interval represented with shaded bands (above), and the median breadth by depth (below).

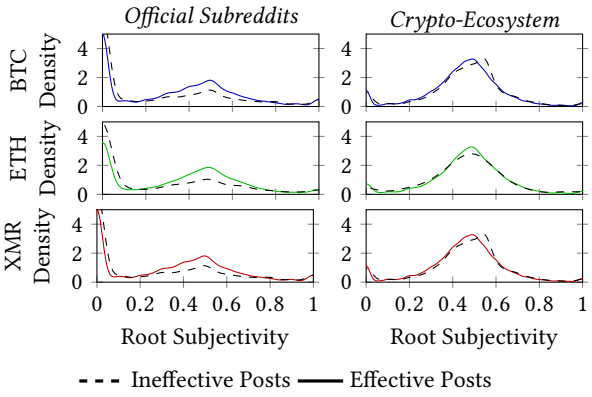


Figure 12: Kernel density estimation plots for *effective posts* that trigger a discussion that includes at least one comment and *ineffective posts* that fail to trigger any subsequent discussion.

to answer the question: *how does the subjectivity of the initial post influence discussion growth?*

First, we compare the subjectivity of effective vs. ineffective discussions at a very high level dichotomy: *did the initial post trigger subsequent discussion or did the discussion consist solely of the initial posts?* We plot the kernel density estimations (KDE) of what we call *effective posts* that initiated discussions of at least one additional comment and *ineffective posts* that failed to initiate a discussion for each of the coins in Figure 12. We see that KDE plots are centered approximately around 0.5, indicating moderate subjectivity, but that ineffective posts tend to skew left, peaking closer to 0.55, for Bitcoin and Monero. Interestingly, ineffective posts about Ethereum appear to remain centered around 0.5 but more evenly across the entire range from 0 to 1, indicating the inclusion of more strongly objective or subjective initial posts.

To gather a more nuanced view of how the initial post in discussions influence the growth, scale, and speed of discussions, we compare normalized mutual information for each of the static discussion properties in Table 3. We also calculated the correlation (Pearson R) of these discussion properties and the subjectivity of the initial post but found significant but low (R<0.05, $p < 0.05$) correlations. Paired with the high normalized mutual information results, we conclude these properties are related but the relationship is not linear. We see that the speed of discussion evolution

Table 3: Normalized Mutual Information between cascade measurements and subjectivity of the root node.

	Official Subreddits			*Crypto-Ecosystem*		
	BTC	ETH	XMR	BTC	ETH	XMR
Virality	0.29	0.36	0.52	0.61	0.68	0.76
Size	0.15	0.21	0.30	0.40	0.49	0.52
# Contributors	0.13	0.18	0.26	0.36	0.43	0.46
Max-Depth	0.10	0.15	0.22	0.29	0.37	0.39
Max-Breadth	0.12	0.17	0.24	0.32	0.40	0.42
Lifetime	0.56	0.68	0.76	0.88	0.92	0.91
Initial Delay	0.37	0.57	0.66	0.75	0.89	0.88

(lifetime and the initial delays) has the greatest mutual information with initial post subjectivity values.

Next, we examined the bivariate kernel density estimations for each of the static discussion cascade measurements and the subjectivity of the initial post (*i.e.,* the "root" of the discussion). We find that the plots within the *Crypto-Ecosystem* dataset illustrate very similar results to the overall distributions of subjectivities shown in Figure 12 at right. Variations in the plots appear to be largely indicative or representative of the differences in the static measurements, highlighted in the previous analyses.

6 SUMMARY

We presented measurement-driven analysis of cryptocurrency discussion spread for three coins of interest on Reddit, and contracted discussion tree patterns in the official coin subreddit and a broader Crypto-Ecosystem on Reddit. Our analysis and novel findings will not only bring the awareness to online discussion spread relevant to cryptocurrencies but will also inform models for forecasting cryptocurrency price that rely on conversations in social media.

We highlight our key findings on *the speed of discussion spread*:

- *Initial Delay* Bitcoin discussions in the official subreddit have the shortest initial delay of approximately 11 minutes. It takes at least 20 minutes for a post to Monero to attract its first comment and over 27 minutes for Ethereum. In the Crypto-Ecosystem delays are larger but coins follow the same trends.
- *Growth Over time* Bitcoin discussions grow the fastest followed by Ethereum then Monero. Bitcoin discussions are the largest at each depth and more efficiently increase the number of comments at each depth than Ethereum and Monero in the official subreddits and Crypto-Ecosystem. Bitcoin shows the fastest growth to largest sizes.
- *Lifetime* Monero discussions have the largest median lifetimes but not the largest maximum lifetimes. Discussions with the largest possible lifetimes are related to the Ethereum cryptocurrency. Interestingly, Bitcoin discussions have the lowest median lifetimes of all three coins.

We highlight our key findings on *the scale of discussion spread*:

- *Volume* On average, Ethereum discussions are smaller than Bitcoin and Monero discussions and Bitcoin discussions are smaller than Monero discussions.
- *Audience Size* The typical discussion is held between 2 and 6 users. In the Crypto-Ecosystem, Bitcoin and Ethereum have

a median of five and four users participating, respectively, while Monero has a median of seven users. ETH < BTC < XMR in terms of typical numbers of participants.

- *Structural properties* Ethereum has the shortest, narrowest, and least viral of all the coins. Monero discussions have the most "viral" of all discussions, on average, with larger maximum depths and breadths. Bitcoin discussions exhibit larger breadths at each of the initial depths.

Future work will extend our preliminary exploration results and further investigate the effect of the original post content on how cryptocurrency discussions spread. More specifically, we will look into content polarity e.g., positive, negative and neutral, novelty, uncertainty and readability. In addition, we will measure user reactions e.g., answer, elaboration, acknowledgement etc. [15] in the discussion cascades influence how discussions spread. Finally, measuring cryptocurrency-related discussion spread before and after the external events of interest e.g., coin price rise, market crash; across a variety of coin types (e.g., popular, stable, scam, etc.); and contrasting signals across multiple social media platforms e.g., Twitter vs. Reddit is of interest.

ACKNOWLEDGMENTS

The research described in this paper was performed at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy. This work was supported by Defense Advanced Research Projects Agency (DARPA) SocialSim program, under agreement 71177. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government. The Reddit datasets were collected by Leidos, the official data provider for the DARPA SocialSim program.

REFERENCES

[1] Lada A Adamic, Thomas M Lento, Eytan Adar, and Pauline C Ng. 2016. Information evolution in social networks. In *Proceedings of the ninth ACM international conference on web search and data mining*. ACM, 473–482.

[2] Lars Backstrom, Jon Kleinberg, Lillian Lee, and Cristian Danescu-Niculescu-Mizil. [n. d.]. Characterizing and curating conversation threads. In *Proceedings of the sixth ACM international conference on Web search and data mining-WSDMâĂŹ13*. 13.

[3] Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 65–74.

[4] Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy* 100, 5 (1992), 992–1026.

[5] Kathleen M Carley. 1999. On the evolution of social and organizational networks. *Research in the Sociology of Organizations* 16, 0 (1999).

[6] Meeyoung Cha, Alan Mislove, Ben Adams, and Krishna P Gummadi. 2008. Characterizing social cascades in flickr. In *Proceedings of the first workshop on Online social networks*. ACM, 13–18.

[7] Meeyoung Cha, Alan Mislove, and Krishna P Gummadi. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the 18th international conference on World wide web*. ACM, 721–730.

[8] Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 925–936.

[9] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. 2016. Do cascades recur?. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 671–681.

[10] Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How community feedback shapes user behavior. *arXiv preprint arXiv:1405.1429* (2014).

[11] Justin Cheng, Jon Kleinberg, Jure Leskovec, David Liben-Nowell, Bogdan State, Karthik Subbian, and Lada Adamic. 2018. Do Diffusion Protocols Govern Cascade Growth? *arXiv preprint arXiv:1805.07368* (2018).

[12] Daejin Choi, Jinyoung Han, Taejoong Chung, Yong-Yeol Ahn, Byung-Gon Chun, and Ted Taekyoung Kwon. 2015. Characterizing conversation patterns in Reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*. ACM, 233–243.

[13] Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K Selcuk Candan, Lexing Xie, Aisling Kelliher, et al. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? *Icwsm* 10 (2010), 34–41.

[14] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: Browsing and voting patterns on Reddit. *IEEE Transactions on Computational Social Systems* 4, 4 (2017), 196–206.

[15] Maria Glenski, Tim Weninger, and Svitlana Volkova. 2018. Identifying and Understanding User Reactions to Deceptive and Trusted Social News Sources. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 176–181. http://aclweb.org/anthology/P18-2029

[16] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.

[17] Sharad Goel, Duncan J Watts, and Daniel G Goldstein. 2012. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*. ACM, 623–638.

[18] Jacob Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters* 12, 3 (2001), 211–223.

[19] Jinyoung Han, Daejin Choi, Jungseock Joo, and Chen-Nee Chuah. 2017. Predicting Popular and Viral Image Cascades in Pinterest. In *ICWSM*. 82–91.

[20] Wenjian Hu, Krishna Kumar Singh, Fanyi Xiao, Jinyoung Han, Chen-Nee Chuah, and Yong Jae Lee. 2018. Who Will Share My Image?: Predicting the Content Diffusion Path in Online Social Networks. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 252–260.

[21] Muhammad Raza Khan. 2017. Cascading Behavior in Yelp Reviews. *arXiv preprint arXiv:1712.00903* (2017).

[22] Young Bin Kim, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang, and Chang Hun Kim. 2016. Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one* 11, 8 (2016), e0161197.

[23] Siddharth Krishnan, Patrick Butler, Ravi Tandon, Jure Leskovec, and Naren Ramakrishnan. 2016. Seeing the forest for the trees: new approaches to forecasting cascades. In *Proceedings of the 8th ACM Conference on Web Science*. ACM, 249–258.

[24] Srijan Kumar, Justin Cheng, Jure Leskovec, and VS Subrahmanian. 2017. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 857–866.

[25] Andrey Kupavskii, Liudmila Ostroumova, Alexey Umnov, Svyatoslav Usachev, Pavel Serdyukov, Gleb Gusev, and Andrey Kustarev. 2012. Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2335–2338.

[26] Kristina Lerman and Rumi Ghosh. 2010. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks. *Icwsm* 10 (2010), 90–97.

[27] Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, and Matthew Hurst. 2009. Information propagation and network evolution on the web. *DA Project, Machine Learning Dept. Carnegie Mellon University* (2009).

[28] Mei Li, Xiang Wang, Kai Gao, and Shanshan Zhang. 2017. A survey on information diffusion in online social networks: Models and methods. *Information* 8, 4 (2017), 118.

[29] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[30] Ross C Phillips and Denise Gorse. 2017. Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on*. IEEE, 1–7.

[31] Eldar Sadikov and Maria Montserrat Medina Martinez. 2009. Information propagation on Twitter. *CS322 project report* (2009).

[32] Eldar Sadikov, Montserrat Medina, Jure Leskovec, and Hector Garcia-Molina. 2011. Correcting for missing data in information cascades. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 55–64.

[33] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.

[34] Senzhang Wang, Xia Hu, Philip S Yu, and Zhoujun Li. 2014. MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1246–1255.

[35] Sha Wang and Jean-Philippe Vergne. 2017. Buzz factor or innovation potential: What explains cryptocurrencies' returns? *PloS one* 12, 1 (2017), e0169556.

[36] Honglin Yu, Lexing Xie, Scott Sanner, et al. 2015. The Lifecyle of a Youtube Video: Phases, Content and Popularity.. In *ICWSM*. 533–542.