Abstract—Software developers around the globe have shown tremendous interests in blockchain with more than seven thousand active blockchain software (BCS) projects on Github. Yet, little research has focused on understanding the challenges encountered by the developers of those projects as well as its users. Therefore, the objective of this study is to better understand the primary areas of challenges encountered by the BCS community. Using a Latent Dirichlet Allocation based topic modeling, we identify discussion topics from the two Blockchain related StackExchange sites. We manually investigated the posts belonging to each topic to understand challenges encountered by the developers. The results of our study revealed that while the ratios of posts on BCS development are increasing, the ratios of posts on mining cryptocurrencies are decreasing. Due to the scarcity of expert blockchain developers, posts on BCS development are more likely to go either answered or encounter more delays than posts on other topics. Based on our findings, we recommend project maintainers to spend efforts to improve documentations on BCS development as the community lacks supporting materials on that area the most.

Index Terms—blockchain, crypto-currency, StackExchange, bitcoin, ethereum, topic modeling, LDA

I. INTRODUCTION

Recently, cryptocurrencies and its underlying blockchain technology has drawn tremendous attentions from software developers. As of May 2019, more than 5,000 developers are regularly contributing to more than eight thousand blockchain repositories hosted on Github and that number is growing very rapidly. Yet, very little research has focused on Blockchain Software (BCS) development practices or to understand the problems being encountered by the BCS community. By BCS community, we imply the developers of BCS as well as the users of blockchain software.

To better assist the blockchain community, it is imperative to understand their interests and difficulties in terms of the blockchain topics they often encounter when developing BCS or using contemporary blockchain tools. Such understanding not only can help the blockchain community but also education, development and research communities that support these emerging ecosystem to better decide when and where to focus their efforts. Without such understanding, developers may not prepare themselves for similar difficulties, educators may develop the wrong educational material and researchers may make incorrect assumptions. Therefore, the primary objective of this study is to understand the primary areas of challenges encountered by the blockchain software community.

On this goal, we mine the StackExchange posts related to blockchain to extract actionable insights from the discussions and closely observe the trends of the topics, the interaction patterns between the topics, and how the trends are influenced by external factors. We leverage StackExchange, since it is one of the most popular online Q&A sites where people ask and answer technical questions. The StackExchange network currently consists of 133 Q&A communities including the well known StackOverflow. Since StackExchange has become the go-to place for the developers to share knowledge and solve issues, we believe mining StackExchange posts can provide valuable insights to understand the contemporary challenges of the blockchain community.

This study utilizes two blockchain related StackExchange sites as dataset sources. The two Blockchain related StackExchange(referred as BSE hereinafter) sites in our study are named after two popular cryptocurrencies: i) bitcoin.stackexchange.com is named after Bitcoin, the first cryptocurrency, which currently has the highest market capitalization and ii) ethereum.stackexchange.com is named after Ethereum, the cryptocurrency with the second highest market capitalization, which aims to facilitate smart-contracts and distributed apps on top of blockchain. We believe that the two BSE sites are ideal data sources for this study, since when a blockchain user encounters a technical challenge s/he is more likely to post a question or conduct a search on one of these sites.

Using a mixed research method, we train a Latent Dirichlet Allocation (LDA) based model to identify the discussion topics from the two BSE sites. LDA has been successfully applied on StackExchange/StackOverflow posts in prior studies to understand developer discussion trends, mobile development trends, android testing issues, requirement engineering questions, and security-related questions from developers.

In Summary, the primary contributions of this study are:
• An empirical investigation of the primary issues and challenges encountered by the blockchain community.
• An investigation of the temporal trends regarding the number of questions, the number of active users, and the time to resolve questions among the two BSE sites.

The remainder of the paper is organized as follows. Section II introduces the research questions of this study. Section III describes our research methodology. Section IV presents the results of this study. Section V discusses the implications of the results. Section VI describes the threats to validity of our findings. Finally, Section VII concludes the paper.

II. RESEARCH QUESTIONS
The primary objective of this study is to better understand the primary areas of challenges encountered by the BCS community. We investigate this primary objective based on three specific research questions. Following subsections present our research questions with a brief rationale behind each question.

A. Discussion Topics
Blockchain users use the two BSE sites to solve their practical problems. Identifying the main discussion topics on those sites can help us identify the major challenges faced by those users. Also, an investigation of the temporal trends of those topics may indicate which topics were challenging during a particular time period and how those challenges evolved. An identification of current challenging areas can help researchers focus on the most pressing areas of interest. Hence, we investigate:

RQ1: What are the main discussion topics related to blockchain and how do users' interests on those topics change over time?

B. Unanswered Questions
We define ‘Unanswered ratio’ as the percentage of questions that did not receive any answer. Due to the increasing difficulties of questions, it is plausible that ‘Unanswered ratios’ may go up on the BSE sites. However, the growing interests among more expert users over the years may increase the likelihood of an answer and therefore decrease ‘Unanswered ratio’. Our next question aims to find out:

RQ2: How do ‘Unanswered ratios’ change over time on the two BSE sites?

C. Answer Intervals
We define the ‘Answer interval’ of a question as the time between a question’s posting and its first answer. Since blockchain is a nascent technology, the number and quality of technical documentations as well as the number of expert users were very limited during the earlier years. Therefore, answer interval may be higher during those years. On the contrary, as only unsolved questions are allowed on StackExchange sites, the earlier questions, which were more basic may have quicker resolution time. However, as the difficulty of the questions grew, answering those questions would require more research and therefore answer intervals may increase. Therefore, our next research question aims to investigate this puzzle:

RQ3: How do answer intervals on the two BSE sites change over time?

III. RESEARCH METHODS
Following subsections describes our research method to answer the three research questions introduces in Section II.

A. Dataset
For this study, we have downloaded December 2018 versions of the two BSE data dumps [2]. Our analyses use three of the seven XML files (i.e., posts.xml, comments.xml, and posthistory.xml). Using SEA-LDA [3], we import the dataset into a MySQL database. Table I shows an overview of our datasets. Between the two BSEs, Bitcoin is the oldest site starting more than 7 years ago and have the highest number of questions posted (around 22K). Ethereum BSE site is less than three years old, however, it has attracted very high level of interests and amassed more than 23K questions within a short period.

B. Data Analysis
We adopted following approaches to compute the measures required to answer our research questions.

1) Answer interval: The posttypeid column in the Posts table indicates whether a post is a question or an answer. For example, posttypeid=1 indicates a question and posttypeid=2 indicates an answer to a question [1]. Using the parrentid column of an answer, we map each answer to the question. Using the creationdate, we compute answer interval, the time between a question’s posting and its first answer.

2) Unanswered question: Using the parrentid field in the Posts table, we identify the questions that did not receive any answer to compute unanswered questions. Using a question’s creationdate we compute unanswered ratio (i.e., percent of questions that went unanswered in each calendar month).

3) Number of active users per month: We use the UserId field in both post_history and comment tables to determine if an user was active during a particular calendar month to compute the number of unique active users for each month.

4) Discussion topics: We queried the Posts table to extract the body field that contains the description of all posts (i.e., both questions and answers). Using the SEA-LDA, we varied the topic_count from 6 to 20 to identify topic models. We evaluated the topic models, using ‘topic coherence’, which measures whether the words in a topic tend to co-occur.

A LDA tool for Software Engineering dataset. Available at: https://github.com/amiangshu/SEA-LDA

<table>
<thead>
<tr>
<th>BSE site</th>
<th># of posts</th>
<th>Time interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Questions</td>
<td>Answers</td>
</tr>
<tr>
<td>Bitcoin</td>
<td>22,130</td>
<td>30,168</td>
</tr>
<tr>
<td>Ethereum</td>
<td>23,817</td>
<td>26,324</td>
</tr>
<tr>
<td></td>
<td>August, 2011 to November, 2018</td>
<td>January, 2016 to November, 2018</td>
</tr>
</tbody>
</table>
together. Among the existing ‘topic coherence’ measures, we use the $C_V$ measure, since it was empirically found to be the most correlated with human interpretability \cite{12}. For the Bitcoin dataset, we found that once topic_count reaches 10, $C_V$ scores do not improve much with additional topics. Hence, we set topic_count=10 for Bitcoin. Similarly, we set topic_count=9 for Ethereum. Although, there is no recommended value for coherence scores, we set $C_V > 0.6$ as the target score for this study as prior study considered this score as good \cite{15}. Our final topic models achieved $C_V=0.63$ score for Bitcoin and $C_V=0.65$ for Ethereum. Following the guidelines from Agarwal et al. \cite{4}, we used only the top ten words from each topic. Using the topics identified, we manually investigate 15 questions from each topic to understand the discussion areas of each topic. During this manual investigation, two of the authors independently assign a discussion category to each topic. For some of the topics that had diverse discussion areas, we considered the areas for the majority of questions. Later, we had discussion sessions to come up with a final category.

IV. Results

In the following subsections, we present the results of our analyses for the three research questions introduced in Section II.

A. Discussion Topics

Our analysis discovered 10 topics in Bitcoin and 9 topics in Ethereum. The topic coherence score ($C_V$) was computed as 0.63 for Bitcoin and 0.65 for Ethereum. In the following subsections, we describe the most common topics and their trends for the two BSE sites. We compute the trend of a topic by computing the percentage of posts on that topic to the total number of questions posted in a given month.

1) Bitcoin topics and trends: Figure 1 shows the trends of the ten identified topics from the Bitcoin BSE from August 2011 to November 2018. Questions related to Exchange and Fund transfer top the list of the most popular topics. Installation and configuration of software from Bitcoin ecosystem comes next. Overall approximately 17% posts are related to coding or understanding blockchain fundamentals. Therefore, majority of the posts on the Bitcoin BSE are not directly related to developing blockchain software. While the ratios of posts on Exchange and Mining are declining, the ratio of questions on Network, Transaction data, and Coding has been increasing during the last three years. Interestingly, the ratio of questions on Fund transfer shows a pattern similar to Bitcoin price, which suggests the price of Bitcoin influencing those questions. While the ratio of questions on Installation, Fees and Fundamentals remain steady. These trends suggest that the ratio of blockchain software developers (i.e., who posts questions on coding, transaction parsing, or network) to blockchain users (i.e., who posts questions on exchange, Keys, or fund transfer) is increasing on the Bitcoin BSE.

Finding 1: Questions related to Exchange, Fund transfer, and Installation were among the most discussed topics at Bitcoin BSE. The ratios of posts on blockchain software development are increasing on the Bitcoin BSE.

2) Ethereum topics and trends: Figure 2 shows the trends of the nine Ethereum topics. While the ratios of posts on network and mining are declining, posts on BCS development (i.e., Solidity, Web apps, and Truffle IDE) are increasing, which is similar to the Bitcoin BSE. We found six overlapping topics (i.e., Fund transfer, Mining, Network, and Keys) between the two BSE sites. While majority of the posts in Bitcoin BSE were not directly related to blockchain software development, approximately 47% posts (i.e., posts belonging to Smart contract, Solidity, Truffle IDE, Web apps, and ERC20 token) in the Ethereum BSE are on solving software development issues. Between the two BCS communities, Ethereum is more developer oriented with higher ratios of posts on BCS development topics. This may be due to the smart-contract\cite{3} programming feature offered by Ethereum, which allows the execution of distributed application on the Ethereum blockchain.

Finding 2: Similar to the Bitcoin BSE, Ethereum BSE also shows increasing ratios of posts on BCS development. On the other hand, the ratios of posts on Mining is also declining on the Ethereum BSE.

B. Unanswered Questions

Figure 3(a) shows the unanswered ratios for the two BSE sites. We observed that during the lifetime of the BSE sites, the unanswered ratios are consistently increasing. It may be due to the increasing difficulties of the questions, as during the earlier days of a BSE site, users post basic questions to learn the preliminary concepts and features. However, after the basic questions are resolved, new questions became more specific and require deeper understandings of the technology. For example, the first question posted on the Bitcoin BSE is: “What open source miner applications are there? Especially to see how the mining process works.”

Finding 3: Similar to the Bitcoin BSE, Ethereum BSE also shows increasing ratios of posts on BCS development. On the other hand, the ratios of posts on Mining is also declining on the Ethereum BSE.

Basic questions like the above example are certain to get multiple responses. On the other hand, the following example shows a question that was asked in June 2016, has received 7 upvotes\cite{5} but have not received any answer to-date. A difficult question similar to the following, requires in depth knowledge:

3Smart-contracts are self-executing contracts with the terms of the agreement between buyer(s) and seller(s) of transactions written using lines of code instead of a legal language. Smart-contracts permit trusted transactions and agreements to be carried out among different anonymous parties without the need for a central authority, legal system, or external enforcement mechanism.

5An upvote to a question indicates another user’s interest in that question.
about blockchain development and therefore very few users are able to answer it.

“I want to call chainActive.Tip()-nHeight in core.cpp. It is but not possible, I am interested in every idea, https://github.com/LIMXTEC/BitSend/blob/DEV-joshafest/src/core.cpp.”

- Bitcoin(#46171)

We also investigated topic-wise unanswered ratios (Figure 3(b) and Figure 3(c)). On the Bitcoin BSE, questions on Fund transfer, Installation, and Coding were among the most unanswered questions. On the Ethereum BSE, Network, Truffle IDE, Web apps, and ERC20 tokens are the top four topics in terms of unanswered ratios. Therefore, i) installation and configuration of the tools and ii) developing software for the blockchain platforms are the two primary areas where users’ expertise are lagging the most.

Finding 3: Due to increasing question difficulties, unanswered ratios show ever-increasing trends. Installation and BCS development are the two areas where blockchain communities lag expertise the most.

C. Answer Intervals

Since the answer intervals are not normally distributed (found using the Shapiro-Wilk test), we use medians to measure the central tendencies for the answer intervals. Figure 4(a) shows the median answer intervals for the two BSE sites. To understand the reasons behind the spikes in answer intervals, we investigated the monthly active users of the two BSE sites during the same period. However, both number of users per month and median answer intervals are time-series. Without a correction for auto correlation, we may observe spurious relationships [19]. Since the results of Augmented Dickey-Fuller [14] tests using the tsseries-R package suggest those time series are indeed auto-correlated, we followed the suggestions of Farnum and Stanton [10] and introduced “first differences” [9] for each time-series. For example, if \( U_m \) indicates the number of users in \( m \)th month, then \( \Delta U_m = U_m - U_{m-1} \).

Using the ccf function from the stats-R package, we compute cross-correlations of number of users, \( \Delta U_m \) with first differences of median answer intervals \( \Delta I_m \). The results suggest that median answer intervals are significantly correlated with the number of active users at lag 0 for Bitcoin (Figure 4(b)). While we do see a negative correlation for Ethereum (Figure 4(c)) at lag 0, it is not statistically significant. Therefore, the number of active users immediately decreases the answer intervals of posted questions on the Bitcoin BSE.

We also investigated topic-wise median answer intervals for the two BSE sites. On the Bitcoin BSE (Figure 5(a)), ques-
tions on Installation, and Coding had higher answer intervals. We also noticed these two topic with the highest unanswered ratios (Section IV-B). On the Ethereum BSE (Figure 5(b)), questions on Network, Web apps had the most delays. Similarly, these two topics also higher unanswered ratios (Section IV-B) as well. These results further validate our conjecture that expertise are lacking in these areas.

Finding 4: The number of active users decreases the median unanswered intervals. Topics with the highest unanswered ratios have higher answer intervals as well, which suggest lack of community expertise on those areas.

V. IMPLICATIONS

In the following subsections, we discuss the implications and recommendations based on our findings.

Discussion Topics

We find the topics regarding Bitcoin and Ethereum that are frequently discussed and the trends of discussions. During the earlier days of the BSE sites, most of the questions were on mining, exchange, keys, and fund transfer. Most of those questions were more likely from the users of cryptocurrencies. However, due to the drop in cryptocurrency prices, the ratio of posts on mining is declining rapidly. However, the ratios of posts on blockchain development areas are increasing, which suggest that the BCS communities interests are moving from cryptocurrencies to BCS development.

A. Unanswered Questions

We observed that unanswered ratios are consistently on upward trends for the two BSE sites. While prior study found unanswered ratio less than 20% for majority of the StackOverflow categories [7], we observed unanswered ratio of more than 20% on each of the two BSE sites. Moreover, during periods of high user interests, unanswered ratio was more than 30% on both BSEs. This result may indicate that despite the growing popularity and hype on blockchain, the number and quality of references and documentations related to blockchain and cryptocurrencies are inadequate. Therefore, a large number of questions remain unanswered. Moreover, the higher ratios of unanswered questions on BCS development areas suggest that compared expert BCS users, the community lacks experts on BCS development.

B. Recommendations

Despite the increasing ratios of posts on BCS development, the number of expert developers to answer those questions may not be growing. As a result posts on development areas are encountering not only higher unanswered ratios but also delayed answers. These may be due to the scarcity of supporting tools and documentation on BCS development [8]. Therefore, the BCS community needs to focus on creating documentation and tutorials on areas indicating growing user interests. For the Ethereum community the focus areas are Truffle IDE, Solidity, Web apps, coding and smart-contracts, and ERC20 tokens. On the other hand for the Bitcoin community the areas are coding, Network, and parsing transaction data.
VI. THREATS TO VALIDITY

Our study was performed only on Bitcoin and Ethereum sub-domains of the StackExchange. It is possible that other domains (i.e., StackOverflow) or other Q&A sites (e.g., Quora) may also contain cryptocurrency-related posts that may be helpful to understand blockchain trends. However, the two BSE sites are the most popular venues for blockchain-related technical discussions and we believe those discussions adequately represent technical topics and trends related to the two cryptocurrencies in this paper.

For LDA, topic number (K) has to be chosen manually. There is no known solution that can automatically determine the most appropriate parameter \( K \). To alleviate this threat, we empirically evaluated the value of K over the range from 6 to 20 to identify the value of \( K \) with the best coherence score. However, it is possible that for a large corpus such as BSEs', a higher value (i.e., \( K > 20 \)) may have provided a better coherence score. Although, with a higher number of topics, we could have identified more fine-grained discussion areas, we do not believe our analysis fails to identify any of the top ten (or nine for Ethereum) discussion areas, since over multiple independent runs, we found stable models \([4]\), where the identified topics were found to be almost identical.

VII. CONCLUSION

We have analyzed the datasets of the two BSE sites to understand the primary challenges and discussion topics among the blockchain software users. We have found that developers’ interest on cryptocurrency is largely influenced by the price of the corresponding cryptocurrency. Community interests are moving from mining to blockchain software development, as the ratios of posts on programming areas are increasing. However, the number and quality of documentations and API references are not improving in accordance with the developers’ interest. As a result, both the ratio of questions that remain unanswered and time to receive the first answer for a question are continuously increasing. Based on our findings, we recommend project maintainers to spend efforts to improve documentations especially on blockchain development areas.

REFERENCES


