ICIS 2019 Proceedings

Crowds, Social Media and Digital Collaborations

# Reinventing the Wheel: Explaining Question Duplication in Question Answering Communities

Xiaohui Liu
*Shanghai Jiao Tong University*, lillian_1st@sjtu.edu.cn

Yijing Li
*School of Information Systems and Technology Management UNSW Business Schoo*, yijing.li@unsw.edu.au

Fei Liu
*Copenhagen Business School*, afl.digi@cbs.dk

Zhao Cai
*University of Nottingham Ningbo China*, zhao.cai@nottingham.edu.cn

Eric Lim
*UNSW Sydney*, e.t.lim@unsw.edu.au

Liu, Xiaohui; Li, Yijing; Liu, Fei; Cai, Zhao; and Lim, Eric, "Reinventing the Wheel: Explaining Question Duplication in Question Answering Communities" (2019). *ICIS 2019 Proceedings*. 27.
https://aisel.aisnet.org/icis2019/crowds_social/crowds_social/27

# Reinventing the Wheel: Explaining Question Duplication in Question Answering Communities

*Short Paper*

**Xiaohui Liu**
Shanghai Jiao Tong University
Shanghai, China
Lillian_1st@sjtu.edu.cn

**Yijing Li**
UNSW Sydney
Sydney, Australia
yijing.li@unsw.edu.au

**Fei Liu**
Copenhagen Business School
Copenhagen, Denmark
afl.digi@cbs.dk

**Zhao Cai**
University of Nottingham Ningbo China
Ningbo, China
Zhao.cai@nottingham.edu.cn

**Eric T.K.Lim**
UNSW Sydney
Sydney, Australia
e.t.lim@unsw.edu.au

## Abstract

*Duplicate questions are common occurrences in Question Answering Communities (QACs) and impede the development of efficacious problem-solving communities. Yet, there is a dearth of research that has sought to shed light on the mechanisms underlying question duplication. Building on the information adoption model, we advance a research model that posits information quality and source credibility as factors deterring users from asking redundant questions within QACs. Furthermore, considering the question-answer dichotomy intrinsic to QACs, we distinguish the quality and credibility of questions from those of answers as distinctive inhibitors of question duplication. We empirically validate our hypotheses on a leading QAC platform by harnessing a deep learning algorithm to detect duplications on over 9,380,000 question pairs. Results revealed that while the credibility of both questions and answers could alleviate question duplication, visual and actionable elements are more effective in preventing question duplication by boosting the quality of questions and answers respectively.*

**Keywords:** Question-Answering, Information Adoption, Question Duplication

## Introduction

Question Answering Communities (QACs) for coders to source for credible solutions to their software development problems have grown in popularity in recent years and are increasingly becoming a viable alternative to traditional web search for problem solving. For example, Stack Overflow, a well-known QAC, has amassed over 10 million registered users as of January 2019 and more than 18 million questions up until mid-2018 (Wikipedia). Before posting their questions, answer seekers often employ search engines to ascertain if similar questions have already been answered (Zhang et al. 2015). If answers to similar questions do exist, answer seekers can directly glean this knowledge to solve their problems without having to post their own. Conversely, answer seekers who fail to find similar questions or locate desired solutions have to post new questions for others to answer. It is hence inevitable for answer seekers to occasionally post *duplicate questions* in that new questions are posted despite the existence of similar ones in QACs (Zhang et al. 2018). Duplicate questions can pose several challenges to the management and operation of QACs (Zhang et al. 2018). Archiving duplicate questions occupies additional storage space and slows down the operation of QACs. Detecting and handling duplicate questions also require more labor cost and increase the technical complexity of the system (Liu et al. 2008). Moreover, duplicate questions attract redundant answers, thus lead to inefficient allocation of scarce attention resources. Finally, duplicate questions serve to disperse answers to a similar question and prolong the delay for newly posted questions to be answered. Answer seekers frustrated by these issues may choose to spam the same question repeatedly, creating a vicious cycle that exacerbates question duplication (Eppler and Mengis 2010).

Acknowledging the prevalence of question duplication in QACs, past studies have centered on the advancement of algorithms for detecting duplicate questions (Zhang et al. 2015). Nevertheless, there have been limited scholarly attempts to comprehend the factors inhibiting the posting of duplicate questions in QACs. To bridge the abovementioned knowledge gap, this study draws on information adoption model and the question-answer dichotomy to identify factors that can mitigate question duplication in QACs (Baltadzhieva and Chrupała 2015; Cheung et al. 2008; Jin et al. 2016). Since duplication stems from non-adoption of existing answers, we adopt information adoption model to elicit information quality and credibility as inhibitors of question duplication (Sussman and Siegal 2003). Considering the unique contextual flavor of QACs, we extends information adoption model by incorporating the question-answer dichotomy, which holds that questions and answers possess distinctive characteristics and require different considerations when processing (Shah et al. 2009). In the context of QACs, *discoverability* is a key characteristic of questions (Harper et al. 2008). Discoverability determines if an answer seeker can discern whether a question resembles his/her own problem. On the other hand, since it is common for each question to have multiple answers, *distinguishability* is the key characteristic of answers (Harper et al. 2008). Distinguishability determines whether an answer seeker can distinguish the viable answer among all the other alternatives.

Information quality in terms of organization and conciseness has been regarded as a focal dimension of information discoverability and distinguishability (Agichtein et al. 2008). A handful of studies have pointed out that QACs with quality questions (Baltadzhieva and Chrupala 2015) and answers (Harper et al. 2008) are less likely to be plagued by question duplication. High quality information usually stands out more and is easier to locate (Gazan 2011). Extrapolated to the context of QACs, high quality questions are more discoverable (Ravi et al. 2014) because they are more likely to attract good answers and become more prominent. On the other hand, answer quality can serve as a differentiating indicator for answer seekers to gauge if an answer addresses their problems. Conceivably, both question and answer quality can make it more likely for answer seekers to solve their problems by adopting existing answers in the QAC, thereby diminishing the probability of question duplication.

Information credibility can function as a heuristic cue for answer seekers to select questions and answers (Alrubaian et al. 2017). Similar to information quality, information credibility aids in mitigating the uncertainty related to the relevance of questions as well as the viability of answers (Albarracín and Vargas 2010). As a social forum, each member in a QAC can earn points for posting questions and answering questions. Each member's accumulated points are displayed as an indicator of his/her reputation (Hoogeveen 2018). Poster's reputations act as an indicator of question credibility and answer credibility, which assist in answer seekers' evaluation of the relevance of questions and the viability of answers (Zha et al. 2018). Consequently, questions posted by or answered by reputable members are less likely to be repeated posted.

Synthesizing extent literature on information adoption model and the question-answer dichotomy, this study endeavors to solicit inhibitors of question duplication in QACs. Specifically, we strive to answer the research question: *what are the inhibitors of question duplication in QACs?* We contextualize our empirical investigation in QAC for coding issues and collected archived questions and answers from a major coding QAC. To quantify the extent of question duplication, we employed a deep learning algorithm based on convolutional neural networks to cluster duplicate questions. Doing so allows us to determine the number of existing duplications for each question. We then attempt to resolve our research question by computing the quality and credibility of each question together with its answers and examine how these focal factors can help to alleviate question duplication. Our preliminary findings are then reported.

## Theoretical Foundation and Hypotheses Foundation

Question duplication can be considered as a consequence of the non-adoption of existing answers, hence is expected to be alleviated if it is made easier for answer seekers to find similar questions and assess if available answers address their problems. Leveraging on Sussman and Siegal (2003)'s information adoption model, we investigate the impact of information quality and source credibility on question duplication. Additionally, information quality bears different implications in QAC. That is, high quality questions are more discoverable whereas high quality answers are more distinguishable. Therefore, we extend the information adoption model by incorporating the question-answer dichotomy, a key contextual factor in QAC. We then construct a research model (see Figure 1) to elicit question quality, answer quality, question credibility, and answer credibility as inhibitors for question duplication.

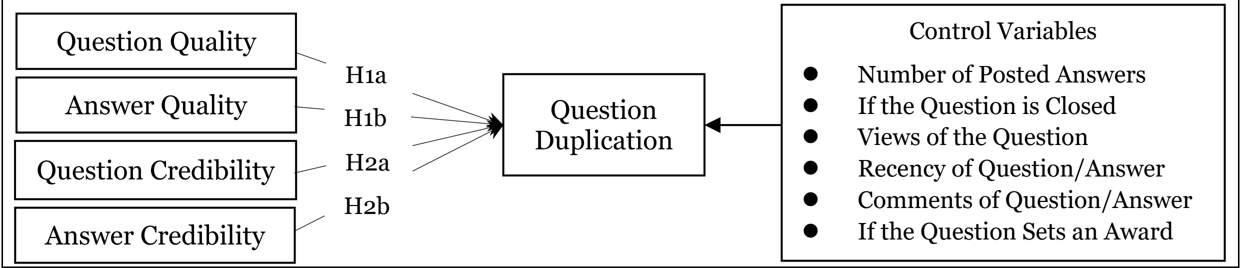
Question Quality
Answer Quality
Question Credibility
Answer Credibility
H1a
H1b
H2a
H2b
Question Duplication
Control Variables
● Number of Posted Answers
● If the Question is Closed
● Views of the Question
● Recency of Question/Answer
● Comments of Question/Answer
● If the Question Sets an Award

**Figure 1. Theoretical Model of the Factors on Question Duplication**

### *Question Answering Community and Information Adoption*

QAC acts as a platform for members to exchange and share knowledge in the form of question answering. Comparing to seeking answers through web search engines, QAC enables answer seekers to solve more specialized and complex problems in a timely fashion (Hoogeveen 2018). Given the extensive adoption of QAC across various domains in recent years, more and more studies have recognized question duplication as a prominent issue (Rodrigues et al. 2017; Zhang et al. 2018). The surging number of duplicated questions have begun to jeopardize the service quality of QAC and in turn exert a detrimental impact on members' engagement and satisfaction. Specifically, for answer seekers, the questions they posted in QAC are often buried under duplicate questions hence elongating the delay for receiving answers. The existence of duplicate questions also scatters potentially viable answers across multiple similar questions, hence making it more challenging to locate the satisfactory answer. For community managers, duplicate questions also pose a challenge since they can clutter up display space and storage space. Researchers and engineers are hence motivated to develop algorithms to automatically detect and remove duplicate questions. For example, Zhang et al. (2015) proposed an automated approach for detecting duplicate questions by calculating similarity in titles, descriptions, latent topics, and tags. Zhang et al. (2018) leveraged on deep learning techniques to capture semantic similarities among questions and postulated a duplication detecting model targeting QACs for coding issues. Moreover, Kaggle, a crowdsourcing community of data scientists and coders that is famous for offering lucrative monetary rewards to problem solvers, posted a machine learning competition for detecting duplicate questions on Quora, a well-known QAC. This competition helps to demonstrate that the question duplication is becoming an increasingly prevalent issue that QACs seek to address. However, there is a dearth of studies focusing on drivers of question duplication. This study posits that question duplication can be alleviated if answer seekers can adopt existing questions and answers to resolve their own problems. For this reason, this study seeks to identify factors that determine if answer seekers can discover and differentiate viable solutions to their problems in a QAC. These factors are hence expected to diminish question duplication.

Sussman and Siegal (2003) drew on elaboration likelihood model to testify the effectiveness of both argument quality and source credibility in facilitating information adoption. These two factors have been leveraged to predict the persuasiveness of information cues in a variety of online communities (Erkan and

Evans 2016). In particular, Cheung et al. (2008) have confirmed the facilitating role of information quality and credibility in affecting consumers' purchase intention. Zha et al. (2018) explored how both quality and credibility dimensions can be utilized to capture information-task fit in social media. Moreover, Jin et al. (2016) empirically investigated the positive impact of both information quality and source credibility on information adoption in healthcare QACs. In this study, we content that the quality and credibility of information available in QACs would also affect if answer seekers can locate desirable answers to adopt. Failing to do so will likely lead to question duplication.

Nonetheless, due to the question-answer dichotomy in QACs, question and answer are two forms of information that play distinctive roles. According to Shah et al. (2009), a typical QAC comprises three components: question posting, answer submission, and a community built around this exchange. Viewed in that light, answer seekers treat existing questions as indices for potentially viable answers (Kim and Oh 2009). Conversely, answers offer solutions that can potentially address answer seekers' problems. Extrapolated in QACs for coding issues, answer seekers rely on question type, difficulty, clarity and topics to gauge if a question shares similarity with their problems (Shah et al. 2009). Upon locating relevant questions, answer seekers then depend on the completeness, accuracy, verifiability, timeliness and supporting sources of each answer to assess which answer can potentially resolve their problems (Shah et al. 2009). As a result, we delineate factors that are associated with the qualities and source credibility of both the questions and the attached answers when investigating their impacts on question duplication.

## Information Quality and Question Duplication

Information quality of questions and answers are distinct concepts in QACs. Specifically, questions with higher quality are expected to clarify the focal issue and demonstrate its relevance (Ravi et al. 2014); Meanwhile, answers of higher quality are usually more comprehensiveness and actionable (Harper et al. 2008). Prior studies have touched upon the role of question quality in online communities (Ravi et al. 2014). For example, Ravi et al. (2014) attested to the connection between topic relevance and question quality by employing a topic modeling approach to predict question quality. Baltadzhieva and Chrupała (2015) have successfully identified constituting dimensions of question quality, including tagging settings, title length, body length, and the presence of code snippets. Past studies have also investigated how answer quality can affect attitudes and behaviors. Shah et al. (2009) highlighted answer quality as an indispensable determinant for the success of QACs and illustrated how the best answer is determined by factors pertaining to the question, the answer itself and the answer giver. If an answer seeker can affirm the relevance of a question to his/her problems and the viability of answers it receives, they are more likely to adopt the solution described in the most promising answer. Therefore, as the quality of questions and the answers they each received improves, answer seekers have a better chance to solve their problems by adhering to existing answers without posting a question of their own. Resultingly, the likelihood for answer seekers to post additional questions plummets, hence alleviating the issue of question duplication.

Specifically, high quality answers with more relevant and well-structured information are more visible to answer seekers (Jin et al. 2016). Consequently, it is more likely for the answer seekers to source for solutions from existing answers rather than posting their own questions. Hence, an increase in question quality can mitigate the issue of question duplication. After identifying a relevant question, answer seekers are likely compelled to look for the most promising solution in multiple answers. For this reason, whether answer seekers can differentiate among multiple answers regarding their viability predicates on the quality of these answers (Raghunathan 1999). Both comprehensiveness and practicality of the answer contribute to a higher answer quality (Fu and Oh 2019). Similarly, answers that are described in greater details are deemed to have higher quality (Cheung et al. 2008). Answers with higher quality appear to be more persuasive for answer seekers thus are more likely to be adopted (Fu and Oh 2019). As a result, despite their distinction, both question quality and answer quality can diminish question duplication. We therefore hypothesize that:

***Hypothesis 1a***: Question quality negatively influences question duplication in a QAC.

***Hypothesis 1b***: Answer quality negatively influences question duplication in a QAC.

## Source Credibility and Question Duplication

Apart from information quality, extant literatures also highlighted source credibility as the other key factor that affects individual's information adoption (Lin et al. 2016). Delving into the past credibility research in online contexts, we define source credibility as the believability of the poster of a question or answer (e.g., poster's reputation). Prior literature has ascertained source credibility as a heuristic indicator for answer seekers to judge whether the information they found is worth adopting. Information disseminated by credible sources are found to be more believable and useful, and are hence more likely to be adopted (Ko et al. 2005; Petty 1986). Hoogeveen (2018) posit reputation as a reflector of source credibility and confirmed the trustworthiness of information from reputable sources.

In QAC, question credibility can help answer seekers gauge the relevance of each question. For instance, a member's reputation can be leveraged by answer seekers to gauge the relevance of the question he/she posted. Answer seekers may hence pay more attention to questions posted by members whose reputation exceeds a certain threshold. Question credibility thereby encourages answer seekers to devote more effort in examining if an existing question resembles their problems. In contrast, answer credibility can help answer seekers differentiate among multiple answers. Since there are often multiple answers existing for the same question, answer seekers can make use of answers' reputation in a comparative fashion to isolate more reputable answers. Answer credibility can thereby draw answer seekers' attention to reputable answers that are more likely to contain working solutions. As a result, question credibility can diminish question duplication by encouraging answer seekers to source for questions relevant to their problems, whereas answer credibility can achieve the same objective by facilitating answer seekers' differentiation of more promising answers from less promising ones. We therefore hypothesize that:

***Hypothesis 2a***: Question credibility negatively influences question duplication in a QAC.

***Hypothesis 2b***: Answer credibility negatively influences question duplication in a QAC.

## Methodology

We rely on a dataset collected from a leading programming community-based question answering sites to validate the advanced research model. The platform enables programmers to either search for existing solutions or posting new questions with regard to the technical issues they encountered. Specifically, our data consists of all details of posts (questions and answers), comments and users' information (e.g., reputation level) from 2014-04-17 to 2014-10-17.

### *Operationalization of Focal Variables*

We operationalize the four focal variables, including question quality, answer quality, question credibility and answer credibility, based on the descriptive information of questions and answers as well as the peripheral testimony on the askers' and answerers' credibility.

**Question Quality:** the descriptions askers posted in QACs are the key for information seekers to identify the similarity between the posted question and the latters' own conundrums. Considering the information characteristics of programming issues, we employ dummy coding on whether codes (*q_if code*), images (*q_if_image*), and extrinsic links (*q_if_url)* are included in the question description. In addition, since the textual features embodied in questions have been attested as the most reliable quality signals in social question answering site (Harper et al. 2008), we add the number of words in questions' body (*q_body_len*), the number of words in questions' title (*q_title_len*), and the number of tags of the question (*q_num_tags*) as our measurement items of question quality.

**Answer Quality:** while question quality measures the discoverability and relevance identifiability of the posted question, the quality of its answers is dependent on the comprehensiveness and practicality of the provided solutions. Accordingly, we contend the variety of the forms of solutions, such as the average number of words in answers' body (*a_body_len*) and the average of the dummy codes on the existence of codes (*a_if_code*), images (*a_if_image),* and urls (*a_if_url*), to capture the comprehensiveness of the structure of the answers. We also include whether there is an accepted answer (*a_if_accept*) in the answer list to gauge the endorsed practicability of the answers.

**Question Credibility and Answer Credibility**: according to information adoption model, the profile of users can be used to measure credibility and hence possess impact on information seekers' adoption propensity(Alrubaian et al. 2017; Zha et al. 2018). Specifically, the platform authorized ratings on users' performances have been deemed as one of the most prominent indicators of their credibility (Jeon 2013).

The focal QAC platform we investigate operates a holistic reputation system whereby each participant is assigned a reputation score that measures his/her performances on posting and answering questions (e.g, upvotes and downvotes on questions/answers). Accordingly, we employ the reputation score of question's owner (*q_OwnerReputation*) as the measurement of question credibility and calculate the average reputation of all answerers engaged in a focal question (*a_OwnerReputation*) to operationalize answer credibility.

**Other controlled variables:** we control for six properties of the posts of questions and answers that may affect question duplication. First, because the platform would automatically close questions with poor quality or off-topic and subjective contents, reliable answers in these closed postings could be discarded so that solution seekers might be compelled to create new questions. We therefore identify whether a question is closed (*q_if_closed*) as control variable. Second, it is conceivable that common problems will be frequently asked regardless of the quality and credibility of existing postings. Accordingly, we adopt the number of views of the question (*q_view_count*) as a proxy of question popularity and conceive it as our second control variable. Third, due to the timeliness of programming problems (Wang and Strong 1996), insightful postings could be obsoleted when they become outdated. As such, we captured the posting timestamps of both questions (*q_asked_time*) and answers (*a_answered_time*) as control variables. Forth, it is likely that the increase of the number of different answers could induce confounding and jeopardize the retrievability of desired solutions (Bawden and Robinson 2009). We hence control for the number of answers published in a focal question (*num_answer*). Fifth, since the question poster can choose to reward the member who provides the best answer, we need to rule out the possibility that questions with rewards would attract more answers regardless of their quality and credibility. Hence, we include whether a question comes with a reward (*q_if_reward*) as a control variable. Finally, since QACs allow users to comment on questions and answers for further interaction, we control the number of comments a question (*q_comment_count*) and its accepted answer (*accep_comment_count*) received. We also control the average number of comments received by all answers to a question (*a_m_comment_count*).

**Question Duplication:** duplication detection has been widely well-recognized as a key challenge in natural language processing (Rodrigues et al. 2017). Leveraging on deep learning techniques, we constructed a Convolutional Neural Network (CNN) to detect duplicate questions and operationalized the temporally weighted average duplications of each unique question (*Num_Dup*) as our dependent variable. In line with Rodrigues et al. (2017), we employ a hybrid CNN to identify duplicate questions based on question titles. The network was pre-trained on Quora dataset which consists of over 400,000 pairs of question titles. We implemented and trained the neural network using a machine with 8 TITAN XP GPU of which each has 16G dedicated memory. Our dataset contains nearly 5000 questions and answers posted during a six-month period. We firstly prepared about 1.25 million possible question pairs, which are enumerated to detect duplication. We can then obtain a temporal weight for each duplicate question through dividing the time it takes for the duplicate to be posted after the original one by the remaining time in the six-month period since the posting of the original question. For an original question posted 90 days into the six-month period, a duplicate posted 45 days later would be assigned a temporal weight of 0.5. Descriptive statistics of all focal variables are summarized in Table 1. Noteworthy, 73.7% of the questions in our dataset are duplicate, which attests to the prevalence of question duplication.

| **Variables** | **Feature** | **Raw Data Statistics** | | | **Data Statistics after logarithmic transformation on the count variables** | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Min** | **Max** | **Mean** | **Min** | **Max** | **Mean** | **Std. Dev.** |
| ***Question Duplication*** | *Num_Dup* | 0 | 792 | 35 | 0 | 2.784 | 1.003 | 0.000 |
| ***Question Quality*** | *q_body_len* | 6 | 2190 | 127 | 1.763 | 4.394 | 2.876 | 0.166 |
| | *q_num_tags* | 1 | 5 | 2.609 | 1 | 5 | 2.609 | 0.129 |
| | *q_if_code* | 0 | 1 | 0.143 | 0 | 1 | 0.143 | 0.123 |
| | *q_if _image* | 0 | 1 | 0.127 | 0 | 1 | 0.127 | 0.099 |
| | *q_if_url* | 0 | 1 | 0.628 | 0 | 1 | 0.628 | 0.157 |
| ***Answer Quality*** | *a_body_len* | 37 | 15,000 | 938 | 0 | 4.176 | 2.744 | 0.212 |
| | *a_if_code* | 0 | 1 | 0.187 | 0 | 1 | 0.187 | 0.189 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *a_if_image* | 0 | 1 | 0.109 | 0 | 1 | 0.109 | 0.129 |
| | *a_if_url* | 0 | 1 | 0.557 | 0 | 1 | 0.557 | 0.154 |
| | *a_if_accept* | 0 | 1 | 0.163 | 0 | 1 | 0.613 | 0.136 |
| ***Question Credibility*** | *q_OwnerReputation* | 0 | 1,070,446 | 27,769 | 0 | 6.030 | 3.672 | 0.000 |
| ***Answer Credibility*** | *a_OwnerReputation* | 0 | 780,089 | 124,492 | 0 | 5.892 | 4.645 | 0.000 |

**Table 1. Descriptive Statistics of Focal Variables**

### *Hypotheses Testing*

Due to the large number of indicators in our model and the co-linearity in our regression equations, we deemed PLS as the most fitting data analysis method. The indicators we adopted in our model represent different aspects of question and answer quality. Therefore, it is more appropriate to model question and answer quality as composites of relevant variables (i.e., second-order formative constructs). In so doing, we can directly hypothesize and estimate the effects of question and answer quality.

Figure 2 illustrates our estimation results by running Partial Least Square (PLS). Overall, all of the hypotheses posited in our model are confirmed. Specifically, question quality possesses a significant and negative impact on question duplication ($\beta = -0.090, p = 0.000$), supporting H1a. Likewise, answer quality could significantly diminish question duplication ($\beta = -0.062, p = 0.000$), hence confirming H1b. The negative paths between question credibility ($\beta = -0.033, p = 0.009$) suggests that questions published by higher rated questioner could be easier to presume its relevancy and attenuates solution seekers' intention to post redundant questions. H2a is hence supported. Meanwhile, the impact of answer credibility on question duplication, as depicted in H2b, is testified by the negative coefficient ($\beta = -0.042$) and its p-value ($p$ = 0.005). Overall, our model is able to articulate 8.8% of the variance in question duplications. Considering our utilizations of secondary objective data and single-item measures, this $R^2$ score is considered satisfactory (Sun 2013) and confirms the predictive power on our model.

Among formative indicators for question quality, providing codes and images are speculated as an efficient way to describe questions and help other questioners to identify relevant question postings. In the meantime, while tagging the question may bolster its discoverability and hence positively related to question quality, lengthy texts in the title and descriptions of the question seems to have little to none effects on its quality. On the flip side, codes and external references are deemed as the optimal modalities to present answers, whereas the indication of accepted answers seems to not relevant to answer quality. Surprisingly, wordy answer may impede users to retrieve practicable solutions from a post, which is in line with the effect of information overload (Bawden and Robinson 2009). All VIF values for our measures are less than 1.5, eliminating the potential of multicollinearity in our structure model (Diamantopoulos and Siguaw 2006). We also calculate the average of inner-construct item correlations for question/answer quality and other intra-constructs items correlations. The results confirm the discriminant validity of our measurement model.

## Expected Contribution to Theory and Practice

This study aims to expand our comprehending of question duplication in QAC for coding issues. First, in spite of its detrimental impact on member experience and community operation, there is a paucity of past research attempting to uncover the underlying inhibitors of question duplication. This study hence strives to spear head the investigation of question duplication in QACs. Second, by synergizing information adoption model with the question-answer dichotomy (Gazan 2011), this study seeks to explicate how both quality and credibility of questions and their answers contribute to mitigating question duplication in QACs. According to the result of our preliminary analysis, having questions and answers with higher quality and credibility can facilitate answer seekers adoption of existing solutions instead of posting their own questions. Consequently, the likelihood for duplicate questions to be posted is diminished. Last but not least, QAC operators who seek to reduce duplicate questions can draw inspiration from the role on the four inhibitors and particularly the information cues embodied in these inhibitors in diminishing question duplication unveiled in this study. For instance, QACs can prioritize the display of questions and answers with higher rated credibility to boost answer seekers' confidence and attention on the existing queries. Our results on the favorable effects of non-textual forms of information could assist platform practitioners to optimize the presentation format and in turn, discourage answer seekers to post potentially duplicate questions. In particularly, while visual elements (i.e., tags, images, and code snippets) that help to index and summarize a question contribute positively to question quality, actionable elements (i.e., code snippets and URLs) that facilitate the test of solutions are essential for boosting answer quality.

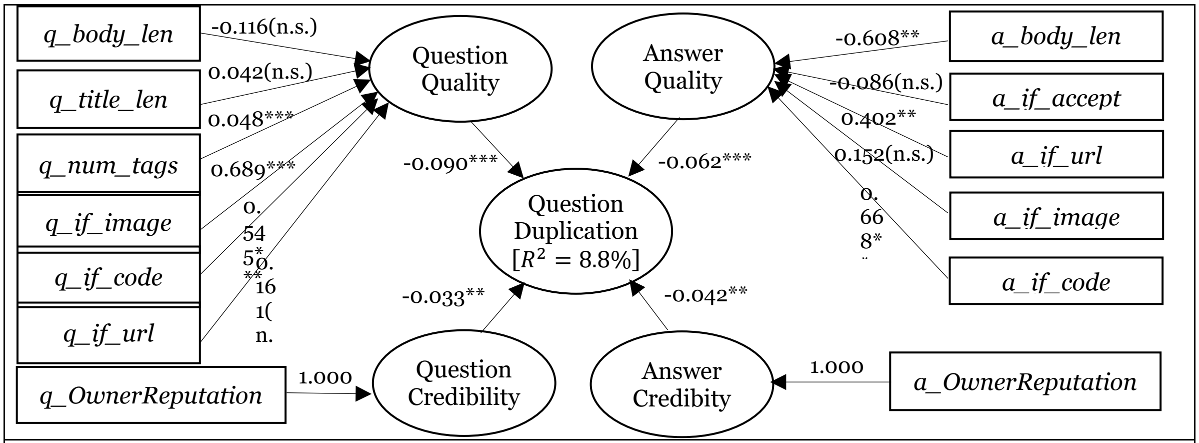

**Figure 2. Model Testing Results**

Note: $* p < 0.05; ** p < 0.01; *** p < 0.001$

## Acknowledgements

Work in this paper was supported by the National Natural Science Foundation of China (NSFC: 71801204)

## References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. 2008. "Finding High-Quality Content in Social Media," *Proceedings of the 2008 international conference on web search and data mining*: ACM, pp. 183-194.

Albarracín, D., and Vargas, P. 2010. "Attitudes and Persuasion," in *Handbook of Social Psychology*. John Wiley & Sons Hoboken, NJ, pp. 394-427.

Alrubaian, M., Al-Qurishi, M., Al-Rakhami, M., Hassan, M. M., and Alamri, A. 2017. "Reputation-Based Credibility Analysis of Twitter Social Network Users," *Concurrency and Computation: Practice and Experience* (29:7), p. e3873.

Baltadzhieva, A., and Chrupała, G. 2015. "Predicting the Quality of Questions on Stackoverflow," *Proceedings of the international conference recent advances in natural language processing*, pp. 32-40.

Bawden, D., and Robinson, L. 2009. "The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies," *Journal of Information Science* (35:2), pp. 180-191.

Cheung, C. M., Lee, M. K., and Rabjohn, N. 2008. "The Impact of Electronic Word-of-Mouth: The Adoption of Online Opinions in Online Customer Communities," *Internet research* (18:3), pp. 229-247.

Diamantopoulos, A., and Siguaw, J. A. 2006. "Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration," *British Journal of Management* (17:4), pp. 263-282.

Eppler, M. J., and Mengis, J. 2010. "The Concept of Information Overload - a Review of Literature from Organization Science, Accounting, Marketing, Mis, and Related Disciplines (2004)," *IEEE Engineering Management Review* (38:1), pp. 3-3.

Erkan, I., and Evans, C. 2016. "The Influence of Ewom in Social Media on Consumers' Purchase Intentions: An Extended Approach to Information Adoption," *Computers in Human Behavior* (61), pp. 47-55.

Fu, H., and Oh, S. 2019. "Quality Assessment of Answers with User-Identified Criteria and Data-Driven Features in Social Q&A," *Information Processing and Management* (56:1), pp. 14-28.

Gazan, R. 2011. "Social Q&A," *Journal of the Association for Information Science and Technology* (62:12), pp. 2301-2312.

Harper, F. M., Raban, D. R., Rafaeli, S., and Konstan, J. A. 2008. "Predictors of Answer Quality in Online Q&a Sites," *human factors in computing systems*, pp. 865-874.

Hoogeveen, D. 2018. "Real and Misflagged Duplicate Question Detection in Community Question-Answering," in: *The School of Computing and Information Systems*. The University of Melbourne.

Jeon, G. Y., Rieh, S. Y. 2013. "Do You Trust Answers?: Credibility Judgments in Social Search Using Social Q&a Sites," *Social Networks* (2:14).

Jin, J., Yan, X., Li, Y., and Li, Y. 2016. "How Users Adopt Healthcare Information: An Empirical Study of an Online Q&a Community," *International Journal of Medical Informatics* (86), pp. 91-103.

Kim, S., and Oh, S. 2009. "Users' Relevance Criteria for Evaluating Answers in a Social Q&a Site," *Journal of the Association for Information Science and Technology* (60:4), pp. 716-727.

Ko, D., Kirsch, L. J., and King, W. R. 2005. "Antecedents of Knowledge Transfer from Consultants to Clients in Enterprise System Implementations," *Management Information Systems Quarterly* (29:1), pp. 59-85.

Lin, X., Spence, P. R., and Lachlan, K. A. 2016. "Social Media and Credibility Indicators: The Effect of Influence Cues," *Computers in Human Behavior* (63), pp. 264-271.

Petty, R. E. 1986. "The Elaboration Likelihood Model of Persuasion," *Advances in Experimental Social Psychology* (19), pp. 123-205.

Raghunathan, S. 1999. "Impact of Information Quality and Decision-Maker Quality on Decision Quality: A Theoretical Model and Simulation Analysis," *decision support systems* (26:4), pp. 275-286.

Ravi, S., Pang, B., Rastogi, V., and Kumar, R. 2014. "Great Question! Question Quality in Community Q&A," *international conference on weblogs and social media*.

Rodrigues, J. A., Saedi, C., Maraev, V., Silva, J. P., and Branco, A. 2017. "Ways of Asking and Replying in Duplicate Question Detection," *joint conference on lexical and computational semantics*, pp. 262-270.

Shah, C., Oh, S., and Oh, J. S. 2009. "Research Agenda for Social Q&a☆," *Library & Information Science Research* (31:4), pp. 205-209.

Sun, H. 2013. "A Longitudinal Study of Herd Behavior in the Adoption and Continued Use of Technology," Mis Quarterly (37:4), pp. págs. 1013-1041.

Sussman, S. W., and Siegal, W. S. 2003. "Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption," Information Systems Research (14:1), pp. 47-65.

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5-33.

Wikipedia. "Stack Overflow." 2019, from https://en.wikipedia.org/wiki/Stack_Overflow

Zha, X. J., Yang, H. J., Yan, Y. L., Liu, K. F., and Huang, C. S. 2018. "Exploring the Effect of Social Media Information Quality, Source Credibility and Reputation on Informational Fit-to-Task: Moderating Role of Focused Immersion," *Computers in Human Behavior* (79), pp. 227-237.

Zhang, W. E., Sheng, Q. Z., Lau, J. H., Abebe, E., and Ruan, W. 2018. "Duplicate Detection in Programming Question Answering Communities," Acm Transactions on Internet Technology (18:3).

Zhang, Y., Lo, D., Xia, X., and Sun, J. 2015. "Multi-Factor Duplicate Question Detection in Stack Overflow," *Journal of Computer Science and Technology* (30:5), pp. 981-997.