Departmental Boundaries and Knowledge Sharing in Corporate Online Communities

Online Appendices

Appendix A. Robustness Checks for the Main Analysis

A.1. Different Model Specifications for the Potential-Dyads Approach

In our potential-dyads approach, we estimated the logit regression model with fixed effects at the provider and seeker levels. To examine the robustness of the estimation results, we apply different model specifications: a linear probability model (LPM) with the provider and seeker fixed effects, as well as the fixed effects on the posting date of the question (Table A.1, Model 1); a logit model with with the provider, seeker, question fixed effects, as well as the fixed effects on the posting date of the question (Table A.1, Model 2); a logit model with the provider and seeker random effects, and fixed effects on the posting date of the question (Table A.1, Model 3); and a logit model with the random effects at the level of both the knowledge provider and seeker, and the two random effects correlated with each other, as well as the fixed effects on the posting date of the question (Table A.1, Model 4). The results are consistent with the results in Table 5.

Table A.1. Results of the Potential-Dyads Approach, with Different Model Specifications

	LPM with provider, seeker, and time fixed effects	Logit with provider, seeker, time, and question fixed effects	Logit with provider and seeker random effects, and time fixed effect	Logit with provider and seeker correlated random effects, and time fixed effect
	(1)	(2)	(3)	(4)
Dependent variable	*Answer*	*Answer*	*Answer*	*Answer*
SameDep	0.002***	0.058***	0.057***	0.054***
	(0.0001)	(0.004)	(0.004)	(0.008)
QLen	0.002***	-	0.042***	0.058***
	(0.0001)		(0.004)	(0.001)
Exp_Ans	0.012***	0.613***	0.706***	0.802***
	(6.77e-5)	(0.004)	(0.003)	(0.032)
Exp_Ques	0.008***	0.151***	0.129***	0.094***
	(0.0002)	(0.003)	(0.003)	(0.030)
PreviousInt	0.003***	0.065***	0.064***	0.083***
	(0.0002)	(0.004)	(0.004)	(0.038)
InterestRele	0.008***	0.845***	0.406***	0.272***
	(0.0003)	(0.027)	(0.011)	(0.091)
OnlineExptSim	0.006***	1.53***	0.682***	0.110***
	(0.0003)	(0.027)	(0.014)	(0.250)
Provider FE	YES	YES	NO/RE	NO/Correlated RE
Seeker FE	YES	YES	NO/RE	NO/Correlated RE

Time FE	YES	YES	YES	YES
Question FE	NO	YES	NO	NO
R^2/Pseudo R^2	0.054	0.210	0.127	-
Specification	LPM	Logit	Logit	MCMC

Notes. The analysis includes 18,670,135 potential knowledge-sharing dyads; knowledge sharing occurs in 687,083 of those dyads. MCMC in model 4 stands for Markov Chain Monte Carlo. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

A.2. Subsample of Employees who Are Constantly Employed During the Study Period

In the main analysis, we construct the potential dyads by considering the employees who have provided one answer and all the questions posted in the study period. This approach makes sure that the analysis considers all the potential dyads. However, it may have considered the dyads where knowledge sharing is unlikely to happen. For example, an employee who resigned during our study period should not be able to answer questions posted after their resignation, and an employee who just joined the company may not be a potential knowledge provider for a question posted a long time ago. To alleviate the concern that the employees who left or just joined the company may express a different knowledge-sharing pattern, we restrict the user sample and focus on a subsample in which employees are constantly employed during the study period. Table A.2 presents the result, which is consistent with that in Table 5.

Table A.2. Results of the Potential-Dyads Approach, with Employees who Are Constantly Employed During the Study Period

	(1)
Dependent Variable	*Answer*
SameDep	0.056***
	(0.004)
QLen	0.057***
	(0.004)
Exp_Ans	0.601***
	(0.004)
Exp_Ques	0.160***
	(0.003)
PreviousInt	0.054***
	(0.004)
InterestRele	0.197***
	(0.011)
OnlineExptSim	-0.103***
	(0.015)
Provider FE	YES
Seeker FE	YES
Time FE	YES
Observations	14,601,758
Pseudo R^2	0.132
Specification	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

A.3. Assumption of the Potential-Ayads Approach

We perform a test at the question level to alleviate the concern about matching each question with all the prospective providers. Specifically, for each question, we calculate the answer ratio of department *d* to question *k*, $AnswerRatio_{dk}$ (i.e., the number of answers from department *d* divided by the number of prospective providers in department *d*) as the dependent variable. We define the independent variables, $SameDep_{dj}$, as whether department *d* is the same as seeker *j*'s department. We aggregate the control variables, *QLen*, *Exp_Ans*, *Exp_Ques*, *PreviousInt*, *InterestRele*, and *OnlineExptSim*, on average at the question-department level. We then examine whether the knowledge seeker's department has a higher ratio of responding to the question. We estimate an OLS regression model with seeker and department fixed effects, and we find that the answer-provider ratio is higher when department *d* is the same as seeker *j*'s department (Table A.3). The results of both analyses are consistent with the main results in Table 5.

Table A.3. Model without an Assumption of the Potential-Dyads Approach

	Question-Department Level
	(1)
Dependent variable	*AnswerRatio*
SameDep	0.002***
	(0.0003)
QLen	-0.008***
	(0.010)
Exp_Ans	0.001***
	(0.0003)
Exp_Ques	0.009***
	(0.002)
PreviousInt	0.010
	(0.010)
InterestRele	0.002***
	(0.0003)
OnlineExptSim	0.005***
	(0.0004)
Seeker FE	YES
Department FE	YES
Time FE	YES
Observations	295,180
R^2/Pseudo R^2	0.170
Specification	OLS

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

A.4. Additional Control Variables

We incorporate ten sets of additional control variables to check the robustness of our potential-dyads approach; the results appear in Table A.4. The first three sets of additional variables pertain to features

of the question text, as supplements to the original controls, $QLen_k$ and $InterestRele_{ik}$. First, we include the topic distribution vector of question k (*Topic1*, *Topic2*, …, *Topic9*; Table A.4, Model 1), which we obtain from the LDA model (see the section that discusses our potential-dyads approach). Second, we construct five custom dictionaries and search for occurrences of the words in question k in each dictionary; we include the counts as additional control variables ($Brand_k$, $Customer_k$, $Administration_k$, $Location_k$, $Product_k$; Table A.4, Model 2). Third, we use the Chinese Linguistic Inquiry and Word Count analyzer (LIWC) to identify the word categories in the question content. We then include the occurrences of the words in question k in the top five content word categories in our question corpus into the model ($CogMech_k$, $Social_k$, $Work_k$, $Affect_k$, and $Time_k$; Table A.4, Model 3).

To describe the question in more detail, we construct the fourth and fifth sets of additional control variables, which are related to the question's difficulty or comprehensiveness. Existing literature suggests that content analysis can measure question quality (Yang et al., 2014). Since we already applied topic modeling for text analysis, we take advantage of this technique to generate topic coverage and topic diversity as proxies for question quality and difficulty. Questions that span multiple or diverse topics tend to promote a more comprehensive understanding, synthesis of information, and critical thinking. This often increases the difficulty of answering due to the need for cross-domain knowledge and the ability to integrate different areas of expertise (Sweller et al., 2011; Xu et al., 2016). First, we include the number of topics the question covers according to the topic vector obtained from the LDA model. For example, if question k's topic vector is $<T_{ik,1}, T_{ik,2}, ..., T_{ik,W}>$, we count the number of non-zero elements as the number of topics of question k. Second, we calculate the diversity of topics in that question based on the LDA topic vector. We use Shannon's entropy, a common measure of diversity in the knowledge-sharing literature (e.g., Kuk 2006): $Diversity_k = -\sum_{w=1}^{W} T_{ik,w} \log T_{ik,w}$.

We then incorporate the subsequent four sets of additional control variables to capture the knowledge-sharing activities of the provider and seeker. First, apart from the provider's experience in the community, we also consider the seeker's previous knowledge-sharing and -seeking behavior. We use $Seeker\_Exp\_Ans_{jk}$ and $Seeker\_Exp\_Ques_{jk}$, to capture the log-transformed sum of seeker j's total number of answers and questions posted before question k (Table A.4, Model 6). Second, in the main analysis, we use $PreviousInt_{ijk}$ to capture the previous interactions from the knowledge seeker responding to the provider. Considering the bidirectional interactions, we include a binary variable $PreviousInt\_ProvidertoSeeker_{ijk}$ to denote whether provider i has previously answered seeker j before

question *k* (Table A.4, Model 7). We also use the discrete number of variables, $PreviousInt\_Num_{ijk}$ and $PreviousInt\_ProvidertoSeeker\_Num_{ijk}$ to consider the intensity of previous interactions (Table A.4, Model 8). Third, utilizing the "best answer" function (labeled by the knowledge seeker; see the screenshot in our research context) in the community, we take account of the high-quality knowledge sharing of the provider and seeker in the past, by counting their best answers before question *k* is posted as $Provider\_BestAns_{ik}$ and $Seeker\_BestAns_{jk}$, respectively (Table A.4, Model 9).

Finally, we include $SameLoc_{ij}$, a binary variable that denotes whether provider *i* and seeker *j* are at the same work location, because employees in the same city may be more inclined to share knowledge with each other (Table A.4, Model 10). The estimations in all models are consistent with those in Table 5.

Table A.4 Results of the Potential-Dyads Approach, with Additional Control Variables

Additional Control Variables	Question Text: LDA vector	Question Text: custom dictionary	Question Text: Chinese LIWC	Question Difficulty : Topic Number	Question Difficulty : Topic Diversity	Seeker's Experience	Previous Interaction : From Seeker to Provider	Previous Interaction : Discrete Number	Best Answers	Same Location
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent variable	*Answer*	*Answer*	*Answer*	*Answer*	*Answer*	*Answer*	*Answer*	*Answer*	*Answer*	*Answer*
SameDep	0.057***	0.057***	0.057***	0.057***	0.057***	0.057***	0.053***	0.047***	0.057***	0.049***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
QLen	0.038***	0.054***	0.031***	0.047***	0.017***	0.042***	0.043***	0.043***	0.041***	0.042***
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Exp_Ans	0.701***	0.701***	0.701***	0.701***	0.693***	0.683***	0.659***	0.563***	0.673***	0.702***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Exp_Ques	0.130***	0.130***	0.130***	0.130***	0.132***	0.133***	0.127***	0.087***	0.110***	0.125***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
PreviousInt	0.065***	0.064***	0.065***	0.065***	0.065***	0.069***	0.058***	-	0.066***	0.066***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)		(0.004)	(0.004)
InterestRele	0.412***	0.421***	0.410***	0.403***	0.588***	0.376***	0.421***	0.472***	0.423***	0.403***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.013)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
OnlineExptSim	0.703***	0.697***	0.702***	0.704***	0.649***	0.882***	0.693***	1.02***	0.780***	0.700***
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.017)	(0.016)	(0.017)	(0.017)	(0.016)
Topic1	-0.015	-	-	-	-	-	-	-	-	-
	(0.013)									
Topic2	-0.006	-	-	-	-	-	-	-	-	-
	(0.015)									
Topic3	0.166***	-	-	-	-	-	-	-	-	-
	(0.015)									
Topic4	0.074***	-	-	-	-	-	-	-	-	-
	(0.019)									
Topic5	-0.023	-	-	-	-	-	-	-	-	-
	(0.015)									
Topic6	0.058***	-	-	-	-	-	-	-	-	-
	(0.018)									
Topic7	-0.048***	-	-	-	-	-	-	-	-	-
	(0.016)									
Topic8	-0.041***	-	-	-	-	-	-	-	-	-
	(0.014)									
Topic9	0.043***	-	-	-	-	-	-	-	-	-
	(0.018)									

Administration	-	-0.063*** (0.005)	-	-	-	-	-	-	-	-
Location	-	-0.122*** (0.008)	-	-	-	-	-	-	-	-
Brand	-	-0.004 (0.004)	-	-	-	-	-	-	-	-
Product	-	-0.060*** (0.003)	-	-	-	-	-	-	-	-
Customer	-	-0.061*** (0.004)	-	-	-	-	-	-	-	-
CogMech	-	-	0.047** (0.014)	-	-	-	-	-	-	-
Social	-	-	0.323*** (0.020)	-	-	-	-	-	-	-
Work	-	-	-0.227*** (0.022)	-	-	-	-	-	-	-
Affect	-	-	0.259*** (0.028)	-	-	-	-	-	-	-
Time	-	-	0.459*** (0.027)	-	-	-	-	-	-	-
QTopicDiversity	-	-	-	-0.009* (0.005)	-	-	-	-	-	-
QTopicNum	-	-	-	-	-0.135*** (0.004)	-	-	-	-	-
Seeker_Exp_Ans	-	-	-	-	-	-0.101*** (0.004)	-	-	-	-
Seeker_Exp_Ques	-	-	-	-	-	-0.013*** (0.003)	-	-	-	-
PreviousInt_ProvidertoSeeker	-	-	-	-	-	-	0.309*** (0.004)	-	-	-
PreviousInt_Num	-	-	-	-	-	-	-	0.389*** (0.002)	-	-
PreviousInt_ProvidertoSeeker_Num	-	-	-	-	-	-	-	0.034*** (0.003)	-	-
Provider_BestAns	-	-	-	-	-	-	-	-	0.106*** (0.004)	-
Seeker_BestAns	-	-	-	-	-	-	-	-	-0.041*** (0.004)	-
SameLoc	-	-	-	-	-	-	-	-	-	0.408*** (0.008)
Provider FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Seeker FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES	YES	YES	YES	YES
Pseudo R^2	0.176	0.176	0.176	0.176	0.176	0.176	0.177	0.182	0.176	0.176
Specification	Logit	Logit	Logit	Logit	Logit	Logit	Logit	Logit	Logit	Logit

Notes. The analysis includes 18,670,135 potential knowledge-sharing dyads; knowledge sharing occurs in 687,083 of those dyads. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Below, we present details for the custom dictionaries and LDA topics. Table A.5 shows explanations of and examples from each dictionary.

Table A.5 Explanations and Examples of Custom Dictionaries

Dictionary	Explanation	Examples
Administration	Words that describe administration levels or departments in the company	Customer-based marketing, Product-based marketing, training

Location	Location-related words, such as city names or the names of geographical regions	Shanghai, East region, Sichuan province
Brand	Words that describe the brands of competing companies and sub-brands that are related to the company in our context	Haidilao, McCormick
Product	Names of company products	Canned food, Soy sauce
Customer	Words that describe customers	The names of key customers

A.5. Text Ming Techniques

A.5.1. LDA

We choose *T* using two methods widely used in the literature (e.g., Puranam et al. 2017; Geva et al. 2019). In the first approach, we maximize the dissimilarity between topics (Cao et al. 2009) by computing a distance between every pair of topics where each is a probability distribution across the vocabulary. The method assumes that LDA performs best when the average cosine distance of topics reaches the minimum. In the second approach, we focus on the goal of deriving topics that differ from one another (Deveaud et al. 2014), and the approach derives the number of topics based on the information divergence (using Jensen-Shannon divergence) between all pairs of topics in a given model; The model assumes that LDA performs best with the number of topics achieving the maximum divergence. We run LDA and calculate the two metrics by varying the number of topics from 5 to 200, as described in Figure A.1. The results from the metric proposed in the first method suggest $T = 5$, and those from the metric proposed in the second method suggests $T = 10$. In the main analysis, we set T to 10. Table A.6 presents the top five keywords and the defined label (summarized by the authors based on keywords) for each topic.

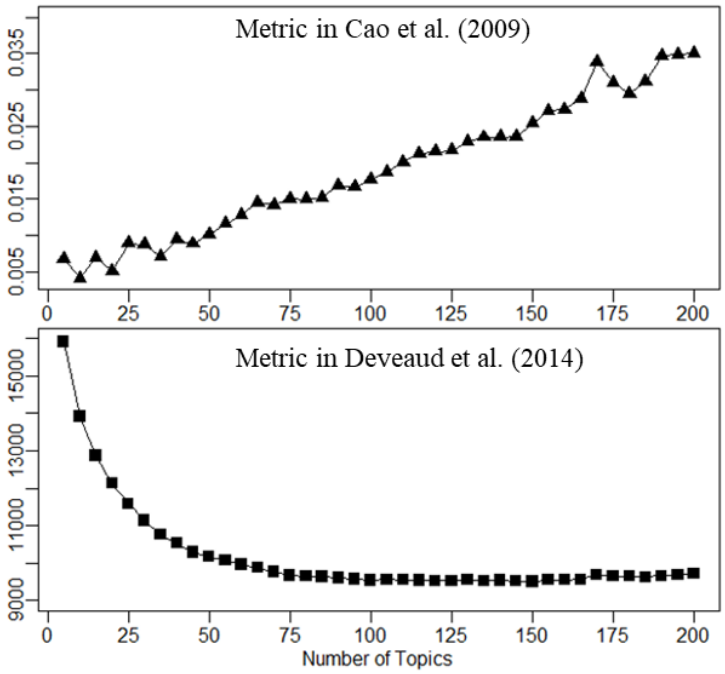

Figure A.1 Metrics in Cao et al. (2009) and Deveaud et al. (2014)

Table A.6 Top Five Keywords and Defined Label of LDA Topics

Topic	Top Five Keywords (Translated from Chinese)	Descriptive Label for the Topic
Topic 1	customer, requirement, product, dishes, sell	understanding demands
Topic 2	dig, focus, attention, interests, attraction	catching interest
Topic 3	chef, quality, cost, price, business	discussing cost and price
Topic 4	situation, problem, encounter, customer, resolve	resolving difficulties
Topic 5	visiting, relationship, time, maintain, shopkeeper	maintaining relationships
Topic 6	distribution, distributor, turnover, account, promote	distribution issues
Topic 7	product A, product B, salty, additive, natural	a product category
Topic 8	product C, product D, product E, sauce, condiments	a different product category
Topic 9	advice, work, effort, opportunities, cheer up	work advice
Topic 10	learn, sharing, platform, ask, training	knowledge sharing

Our main empirical results are not sensitive to the number of topics defined in LDA. We show that if generating control variables with LDA of different numbers of topics, such as $T = 5$, $T = 15$, and $T = 20$, the main results are consistent with that in Table 5.

A.5.2. Word Embedding

We employ an alternative text mining technique, word embedding, a deep-learning-based algorithm (Mikolov et al. 2013; Song et al. 2018) to generate control variables and further validate our results. This method considers contextual information of words and has become increasingly popular in the IS literature (Wu et al. 2021; Yu et al. 2023). We train an embedding model using the skip-gram method in word2vec on the entire Q&A content collection. The implemented word embedding model generates

semantic vectors, representing predicted neighboring words for each word in the corpus. After obtaining word embedding vectors, we then construct the vectors for Q&A content using a composition function by unweighted averaging corresponding to the words in the content. Using the content embedding vectors, we re-calculate the variables $InterestRele_{ik}$ and $OnlineExptSim_{ijk}$ and re-estimate Equation (1).

We use the module *Gensim* in Python to construct the word embedding model. We set the dimension of a word vector to D = 200 (results are consistent with D = 100 and D = 300). This parameter is a default setting suggested by *Gensim* and is commonly used in implementing Chinese word embeddings (Song et al., 2018). Table A.7 presents the estimation results with control variables generated from word embedding, and the result is consistent with that in Table 5.

Table A.7 Results of the Potential-Dyads Approach, with Control Variables Using Word Embedding

	(1)
Dependent Variable	*Answer*
SameDep	0.054***
	(0.004)
QLen	0.008*
	(0.004)
Exp_Ans	0.668***
	(0.004)
Exp_Ques	0.134***
	(0.003)
PreviousInt	0.063***
	(0.004)
InterestRele_WordEMB	0.444***
	(0.007)
OnlineExptSim_WordEMB	1.01***
	(0.018)
Provider FE	YES
Seeker FE	YES
Time FE	YES
Pseudo R2	0.176
Specification	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

A.6. Independence of Department Proximity from Online Expertise Similarity

We conduct two analyses to demonstrate the independence of department proximity from online expertise similarity. First, we assess the effect of *SameDep* in subsamples above and below the median of *OnlineExptSim*. Models 1 and 2 in Table A.8 present the results, indicating that the provider and seeker in the same department, regardless of whether they have a low or high level of online expertise similarity, are more likely to share knowledge.

Second, we generate *OnlineExpSim_Ans* using user-generated answers alone, instead of both answers and questions (*OnlineExpSim*) in the main analyses. A user's answers more accurately reflect

their expertise, whereas questions indicate the knowledge they seek. Model 3 in Table A.8 shows that the coefficient of *SameDep* remains significant. This suggests that the inclination to share knowledge within departmental boundaries is robust to changes in how expertise similarity is defined.

Table A.8 Independence of Department Proximity from Expertise Similarity

	Subsample		Generate using answers
	Above the median	Below the median	
	(1)	(2)	(3)
Dependent Variable	*Answer*	*Answer*	*Answer*
SameDep	0.093***	0.051***	0.057***
	(0.009)	(0.004)	(0.004)
QLen	-0.016**	0.068***	0.035***
	(0.010)	(0.005)	(0.004)
Exp_Ans	0.485***	0.991***	0.782***
	(0.005)	(0.008)	(0.004)
Exp_Ques	0.108***	0.154***	0.096***
	(0.006)	(0.004)	(0.004)
PreviousInt	0.078***	0.050***	0.493***
	(0.010)	(0.004)	(0.011)
InterestRele	0.980***	0.183***	0.057***
	(0.022)	(0.013)	(0.004)
OnlineExptSim	0.946***	-3.40***	-
	(0.020)	(0.387)	
OnlineExptSim_Ans	-	-	0.685***
			(0.015)
Provider FE	YES	YES	YES
Seeker FE	YES	YES	YES
Time FE	YES	YES	YES
Observations	9,330,975	9,292,584	18,670,135
Pseudo R^2	0.261	0.105	0.175
Specification	Logit	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Appendix B. Additional Information for the Same-Name Design

B.1. Illustration of the Same-Name Design

We illustrate the identification challenge and strategy for our empirical design using the same-name feature in Figure B.1. In our context, prospective knowledge providers make knowledge-sharing decisions based on two main information sources: the question text and the knowledge seeker's attributes. If we aim to identify the effect of the seeker's department on the provider's knowledge-sharing probability, then we should exclude the effects of the question text. Although we include control variables regarding the question text in the potential-dyads approach, we cannot control all possible characteristics. As shown in Figure B.1, Panel A, suppose we observe that provider *i* answers question *k* from seeker *j*, and *i* and *j* are in the same department. This outcome can be explained by their same department, other observables (considered in $Control_{ijk}$), and/or unobservable confounders ($Confounders_{ij}$) that we are unable to include in our analyses. For example, perhaps *i* and *j* may use similar terminologies or technical terms because they are in the same department, and similar terminologies or technical terms increase *i*'s probability of answering *j*'s question (the dashed lines).

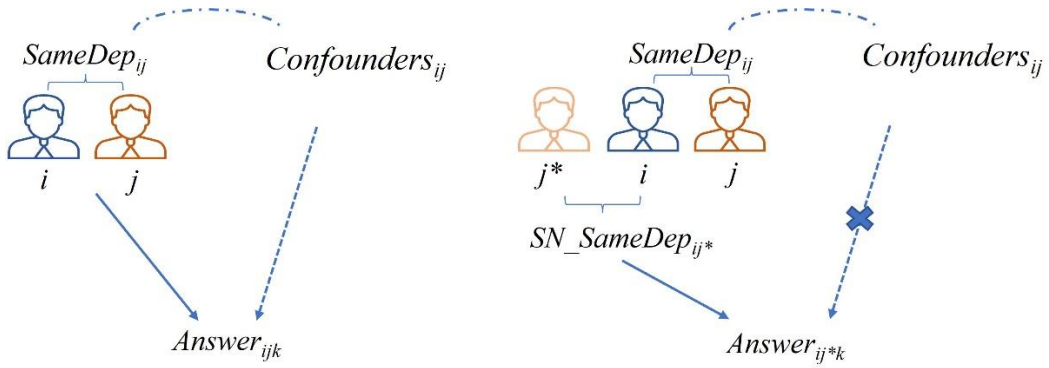

Figure B.1. Identification Challenge and Strategy

B.2. Robustness Checks for the Same-Name Design

To further take into account the role of previous activities in the mistaken-identity scenario, we match the treatment group, where the mistaken-identity scenario can potentially happen, and the control group, where the mistaken-identity scenario does not happen with a set of matching variables describing users' activities on the community. Specifically, we use propensity score matching (PSM) to obtain comparable control and treated groups.

In the matching process, the treatment group is the same-name mistaken-identity scenario can potentially happen, where the seeker's same-name peer is in the same department as the provider, but the seeker is not. The control group presents that the same-name mistaken-identity scenario does not happen, where the seeker's same-name peer is not in the same department as the provider. Using PSM, we match each treated dyad with a most "similar" control dyad in terms of the following characteristics: the number of answers and questions posted by the provider before (*Exp_Ans* and *Exp_Ques*), the number of answers and questions posted by the seeker before (*Exp_Seeker_Ans* and *Exp_Seeker_Ques*), whether the provider has received an answer from the seeker before (*PreviousInt*), and the similarity between the provider's and seeker's online expertise (*InterestRele*).

We first sort all observations in a random order to ensure the ordering does not affect the subsequent matching. Then, we run a logit regression based on the pretreatment variables mentioned earlier and obtain the predicted propensity scores. We match the observations in the treatment and control groups with the most similar propensity scores with the nearest-neighbor matching algorithm. We perform t-tests of equality of means after the matching to check whether our PSM adequately balances characteristics between the treatment and control groups. The results are presented in Table B.1. After matching, across all the matching variables, the differences in mean between the treatment and control groups are not statistically significant, suggesting that matching helps reduce the bias associated with the observable characteristics. Finally, we estimate our same-name models using the new sample created by PSM, and the results are in Table B.2. The findings are consistent with those in Table 5.

Table B.1. Matching Outcome

Variables	Before matching		After matching				
	Mean Treated	Mean Control	Mean Treated	Mean Control	%bias	*t*-Statistics	*p*-value
Exp_Ans	2.9354	3.9001	2.9354	2.8912	1.8	1.24	0.215
Exp_Ques	0.60432	0.88221	0.60432	0.60793	-0.4	-0.27	0.790
Exp_Seeker_Ans	5.0907	5.8491	5.0907	5.0695	1.4	0.99	0.324
Exp_Seeker_Ques	1.8677	3.4838	1.8677	1.8655	0.1	0.15	0.884
PreviousInt	0.08542	0.16627	0.08542	0.0783	2.2	1.78	0.075
InterestRele	0.65447	0.80847	0.65447	0.64952	1.3	0.77	0.440

Table B.2. Results for Knowledge Seekers with the Same-Name Feature, with the Matched Sample

	(1)	(2)	(3)	(4)
Dependent Variable	*Answer*	*Answer*	*Answer*	*Answer*
SN_SameDep	0.305**	0.358**	0.340**	0.392**
	(0.143)	(0.155)	(0.146)	(0.157)
QLen	0.361*	0.335	0.355*	0.327
	(0.200)	(0.209)	(0.205)	(0.214)
Exp_Ans	0.513***	0.405***	0.494***	0.376***
	(0.127)	(0.139)	(0.131)	(0.143)
Exp_Ques	-0.088	-0.100	-0.059	-0.072
	(0.064)	(0.064)	(0.064)	(0.065)
PreviousInt	0.019	0.018	-0.052	-0.046
	(0.112)	(0.114)	(0.116)	(0.118)
InterestRele	-0.278	-0.277	-0.316*	-0.317
	(0.185)	(0.189)	(0.189)	(0.194)
OnlineExptSim	-0.267	-0.191	-0.485	-0.403
	(0.329)	(0.359)	(0.334)	(0.363)
AdditionalControls	NO	YES	NO	YES
Seeker FE	YES	YES	YES	YES
Provider FE	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
Observations	15,913	15,913	15,314	15,314
Pseudo R^2	0.098	0.113	0.097	0.113
Specification	Logit	Logit	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Appendix C. Additional Analyses for Mechanisms

C.1. Mediation Analyses

C.1.1. The Role of Receiving the Best Answers

To provide in-depth insights for knowledge-focused motivations, we conduct mediation analyses for the best answers. In our context, the "best answer" is selected by the knowledge seeker, serving as a proxy for answer helpfulness, which measures how well an answer meets the seeker's needs (Tian et al. 2013). If knowledge-focused motivations play an important role, being in the same department increases the knowledge-sharing possibility because knowledge providers know that doing so will increase the possibility of seekers' recognition, which further motivates more contributions. In this light, users are motivated to share answers when they know their responses could be recognized as the "best," which aligns with our main idea of knowledge-focused motivations.

However, applying a classic mediation model for "best answer" is not feasible because an answer can only be selected as "best" if it has already been contributed (i.e., the dependent variable *Answer* = 1). Additionally, using the cumulative number of best answers at the dyadic level may not accurately reflect the mediating role. We adopt an alternative set of analyses that align with the intuition behind the mediation analyses. First, we assess whether the answer from the provider in the same department as the seeker is more likely to be selected as the "best answer." Second, we evaluate whether a provider with a history of having their answers selected as "best" is more likely to respond to the focal question. Specifically, we estimate the following models:

$$Best_{ijk} = \beta_{10} + \beta_{11} SameDep_{ij} + Control_{ijk} + \delta_i + \theta_j + \gamma_k + \epsilon_{ijk} \quad \text{(C.1)}$$

$$\begin{aligned} Prob\left(Answer_{ijk} = 1 \middle| x\right) = {} & \beta_{20} + \beta_{21} SameDep_{ij} + \beta_{22} PreviousBest\_SameDep_{ijk} \\ & + Control_{ijk} + \delta_i + \theta_j + \gamma_k + \epsilon_{ijk} \end{aligned} \quad \text{(C.2)}$$

where the variable $Best_{ijk}$ represents whether the answer contributed by provider *i* to seeker *j*'s question *k* is selected as the best answer, and the variable $PreviousBest\_SameDep_{ijk}$ is a dummy variable indicating whether provider *i*'s previous answers have been selected as the best answers before seeker *j*'s question *k* is posted. We also replace $PreviousBest\_SameDep_{ijk}$ with $PreviousBestNum\_SameDep_{ijk}$, which counts the number of provider *i*'s best answers before seeker *j*'s question *k*. Table C.1 shows that the coefficients of *SameDep* in Model 1, $PreviousBest\_SameDep_{ijk}$ in Model 2, and $PreviousBestNum\_SameDep_{ijk}$ in Model 3 are all positive and statistically significant. This analysis

supports the expectation that the answer inclination is associated with a higher possibility of a "best answer" badge. Also, we note that after considering the best answers received from the same department, the estimation for $SameDep_{ij}$ is still significant, albeit having a smaller magnitude (vs. that in Model 1), indicating that hoping to receive the best answer badge is only one of the reasons.

Table C.1. The Role of Receiving Best Answers

	(1)	(2)	(3)
Dependent Variable	*Best*	*Answer*	*Answer*
SameDep	0.069*	0.057***	0.057***
	(0.035)	(0.004)	(0.004)
PreviousBest_SameDep	-	0.334***	-
		(0.017)	
PreviousBestNum_SameDep	-	-	0.504***
			(0.020)
QLen	-0.061	0.042***	0.042***
	(0.040)	(0.004)	(0.004)
Exp_Ans	-0.250***	0.701***	0.702***
	(0.028)	(0.004)	(0.004)
Exp_Ques	-0.014	0.126***	0.126***
	(0.027)	(0.003)	(0.003)
PreviousInt	0.129***	0.064***	0.064***
	(0.035)	(0.004)	(0.004)
InterestRele	0.013	0.406***	0.406***
	(0.102)	(0.011)	(0.011)
OnlineExptSim	-0.123	0.710***	0.709***
	(0.132)	(0.016)	(0.016)
Provider FE	YES	YES	YES
Seeker FE	YES	YES	YES
Time FE	YES	YES	YES
Observations	687,083	18,670,135	18,670,135
Pseudo R^2	0.145	0.176	0.176
Specification	Logit	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

C.1.2. The Role of Historical Expertise Similarity

We refine our conceptualization of expertise similarity to segmented online expertise similarity into: (1) historical expertise similarity: based on tacit, department-specific knowledge accumulated over time and difficult to observe directly, and (2) recent expertise similarity: based on explicit, short-term knowledge within observable recent interactions. The distinction isolates historical (tacit) from recent (explicit) dimensions, allowing us to examine the mediation role of measurable tacit common ground based on expertise similarity. Even though, it is important to note that expertise similarity, as measured through online content, likely captures only a small fraction of tacit common ground. A large part of tacit common ground, such as shared culture and experiences within the department, is difficult to transfer into measurable content.

Based on this refinement, we have conducted a three-step mediation analysis (Bai et al. 2020) to examine the role of historical (tacit) expertise similarity in the relationship between departmental boundaries and knowledge-sharing likelihood, while controlling for recent expertise similarity. Specifically, we estimated the following mediation models:

$$Prob(Answer_{ijk} = 1|x) = \beta_{30} + \beta_{31}SameDep_{ij} + \beta_{32}Recent\_ExpSim_{ijk} + Control_{ijk} + \delta_i + \theta_j + \gamma_k + \epsilon_{ijk} \quad \text{(C.3)}$$

$$Historical\_ExpSim_{ijk} = \beta_{40} + \beta_{41}SameDep_{ij} + \beta_{42}Recent\_ExpSim_{ijk} + Control_{ijk} + \delta_i + \theta_j + \gamma_k + \epsilon_{ijk} \quad \text{(C.4)}$$

$$Prob(Answer_{ijk} = 1|x) = \beta_{50} + \beta_{51}SameDep_{ij} + \beta_{52}Historical\_ExpSim_{ijk} + \beta_{53}Recent\_ExpSim_{ijk} + Control_{ijk} + \delta_i + \theta_j + \gamma_k + \epsilon_{ijk} \quad \text{(C.5)}$$

where the variable $Historical\_ExpSim_{ijk}$ measures the expertise similarity between provider *i* and seeker *j* based on their answers prior to the last day *i* contributed answer(s) before question *k* was posted, and the variable $Recent\_ExpSim_{ijk}$ captures the expertise similarity between provider *i* and seeker *j* on the last day *i* contributed answer(s) before question *k* was posted. $Controls_{ijk}$ represents the vector of control variables used in the main analysis, excluding $OnlineExpSim_{ijk}$. All other symbols retain the same meanings as in the main analysis.

Table C.2 shows the results of the mediation analysis. Model 1 evaluates the total effect of the independent variable on the dependent variable with the control of recent expertise similarity, which remain consistent with the main results ($p < 0.001$). The effect of being in the same department on the mediator (historical expertise similarity) in Model 2 and the effect of the mediator on answer probability in Model 3 are both significantly positive. The Sobel-Goodman mediation test reveals that the mediation effect of historical expertise similarity, captured by the product of the above two estimators, is also significant ($p < 0.001$). However, the proportion of the total effect mediated is relatively small at 0.35%. This finding suggests that the part of tacit common ground reflected by historical expertise similarity partially mediates the observed relationship in a small portion, and other factors, such as unmeasured tacit dimensions (e.g., shared culture and experiences) and broader social and relational dynamics, likely explain the majority of the relationship.

Table C.2. The Role of Historical Expertise Similarity

	(1)	(2)	(3)
	Dependent Variable on Independent Variable	**Mediator on Independent Variable**	**Dependent Variable on Mediator & Independent Variable**
Dependent Variable	*Answer*	*Historical_ExpSim*	*Answer*
SameDep	0.002***	0.001***	0.002***
	(0.0001)	(0.0001)	(0.0001)
Historical_ExpSim (Mediator)	-	-	0.006*** (0.0004)
QLen	0.002***	-0.029***	0.002***
	(0.0001)	(0.0001)	(0.0001)
Exp_Ans	0.013***	0.109***	0.012***
	(0.0001)	(0.0000)	(0.0001)
Exp_Ques	0.008***	-0.046***	0.008***
	(0.0001)	(0.0001)	(0.0001)
PreviousInt	0.003***	0.002***	0.003***
	(0.0002)	(0.0001)	(0.0002)
InterestRele	0.011***	0.462***	0.008***
	(0.0003)	(0.0002)	(0.0003)
Recent_ExpSim	0.002***	-0.383***	0.004***
	(0.0005)	(0.0003)	(0.0005)
Provider FE	YES	YES	YES
Seeker FE	YES	YES	YES
Time FE	YES	YES	YES
R-squared	0.054	0.897	0.054
Specification	OLS	OLS	OLS

Notes. The analysis includes 18,670,135 potential knowledge-sharing dyads; knowledge sharing occurs in 687,083 of those dyads. Heteroskedasticity-consistent robust standard errors are in parentheses. To ensure comparability, we use the OLS specification for all three models. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

C.2. Mechanism Analyses Using the Same-Name Design

We have replicated the mechanism analyses for the same-name design. All the findings are consistent with those obtained from the full dataset with *SameDep* as the independent variable.

C.2.1. Knowledge-Focused Motivations

Subsample of question types

We use *SN_SameDep* as the key independent variable and replicate the subsample analysis across different question types with the same-name dataset. The results, presented in Table C.3, show that the coefficients of *SN_SameDep* are positive and significant. The magnitude of the estimates indicates that the inclination to answer in the same-name mistaken scenario for the same department varies across question types, suggesting that knowledge-focused motivations cannot be neglected.

Table C.3. Subsample of Question Types: Same-Name Dataset

Question Type	Product	Customer	General
	(1)	(2)	(3)
Dependent Var.:	*Answer*	*Answer*	*Answer*
SN_SameDep	0.00420*	0.02535***	0.03532***
	(0.00234)	(0.00624)	(0.00625)
QLen	0.00330***	-0.00354***	-0.00129**
	(0.00055)	(0.00101)	(0.00060)
Exp_Ans	0.00212***	0.00098	0.00319***
	(0.00041)	(0.00066)	(0.00044)
Exp_Ques	0.00322***	0.00646***	0.00479***
	(0.00090)	(0.00155)	(0.00101)
PreviousInt	0.00441***	-0.00112	0.00535***
	(0.00109)	(0.00218)	(0.00135)
InterestRele	0.00290*	0.01873***	0.00878***
	(0.00165)	(0.00272)	(0.00189)
OnlineExptSim	0.02409***	0.01002***	0.01981***
	(0.00180)	(0.00294)	(0.00209)
Provider FE	YES	YES	YES
Seeker FE	YES	YES	YES
Time FE	YES	YES	YES
R^2	0.053	0.056	0.058
Observations	627,440	208,725	503,470
Specification	OLS	OLS	OLS

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

The role of job function proximity

We replicate the analysis related to the job function for the same-name dataset. Except for *SN_SameDep as the independent variable, we include SN_DiffDepSameFunc*, which indicates the mistaken scenario where the seeker has a different job function from the provider, and the same-name peers of the seeker are from the different departments but have the same job function with the provider. Table C.4 shows that this same-name mistaken scenario for the same function also has a higher probability of sharing knowledge, indicating the existence of knowledge-focused motivations.

Table C.4. The Role of Job Function Proximity: Same-Name Dataset

	(1)
Dependent Var.:	*Answer*
SN_SameDep	0.099***
	(0.014)
SN_DiffDepSameFunc	0.385***
	(0.056)
QLen	0.023*
	(0.011)
Exp_Ans	0.003
	(0.008)
Exp_Ques	0.039**
	(0.012)
PreviousInt	0.044**
	(0.014)
InterestRele	0.426***
	(0.040)
OnlineExptSim	2.12***
	(0.071)
Provider FE	YES
Seeker FE	YES
Time FE	YES
Pseudo R^2	0.131
Observations	1,168,416
Specification	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

C.2.2 Social-Related Motivations

We replicate the subsample analysis for professional and non-professional questions with the same-name dataset, treating *SN_SameDep* as the key independent variable. Table C.5 presents the results, which suggest positive and statistically significant coefficients for both professional and non-professional questions. This finding indicates that social-related motivations also drive the knowledge-sharing inclination in the same-name mistaken scenario for same-department colleagues.

Table C.5. Subsample of Professional and Non-Professional Questions: Same-Name Dataset

Question Type	Professional	Non-Professional
	(1)	(2)
Dependent Variable	*Answer*	*Answer*
SN_SameDep	0.302***	0.654***
	(0.064)	(0.116)
QLen	-0.012	0.230***
	(0.013)	(0.028)
Exp_Ans	-0.006	0.039*
	(0.009)	(0.018)
Exp_Ques	0.031*	0.087**
	(0.013)	(0.030)
PreviousInt	0.033*	0.085*
	(0.015)	(0.036)
InterestRele	0.525***	0.003
	(0.045)	(0.087)
OnlineExptSim	1.97***	3.31***
	(0.077)	(0.203)
Provider FE	YES	YES
Seeker FE	YES	YES
Time FE	YES	YES
Pseudo R^2	0.130	0.118
Observations	989,449	178,967
Specification	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

C.3. Results for Knowledge-Focused Motivations of the Same-Location Knowledge-Sharing Inclination

Subsample of question types

First, based on the classification of questions (i.e., product-related, customer-related, and general questions), we conduct three subsample analyses to examine the extent of same-location answer inclination in the three types of questions. We present the results in Table C.6. A comparison of the magnitude of the estimations with the three subsamples indicates that questions also matter when knowledge providers use location information to make answer decisions. The results of the t-tests reveal statistically significant differences in the coefficients for *SameLoc* across question types. Specifically, a significant difference exists between product-related and customer-related questions (t = -0.790, p = 0.043), as well as between product-related and general questions (t = -0.377, p = 0.071).

Table C.6. Subsample of Question Types: Same-Location Answer Inclination

Question Type	Product	Customer	General
	(1)	(2)	(3)
Dependent Var.:	*Answer*	*Answer*	*Answer*
SameLoc	0.02084***	0.02440***	0.02283***
	(0.00064)	(0.00125)	(0.00082)
QLen	0.00226***	-0.00116*	0.00120***
	(0.00023)	(0.00047)	(0.00026)
Exp_Ans	0.01258***	0.01119***	0.00733***
	(0.00011)	(0.00016)	(0.00013)
Exp_Ques	0.00712***	0.00844***	0.00665***
	(0.00023)	(0.00037)	(0.00028)
PreviousInt	0.00359***	0.00273***	0.00203***
	(0.00029)	(0.00051)	(0.00036)
InterestRele	0.01048***	-0.00059	0.00887***
	(0.00047)	(0.00081)	(0.00057)
OnlineExptSim	0.00219***	0.01063***	0.01177***
	(0.00048)	(0.00078)	(0.00059)
Provider FE	YES	YES	YES
Seeker FE	YES	YES	YES
Time FE	YES	YES	YES
Pseudo R^2	0.10870	0.10980	0.14022
Observations	7,199,115	2,328,865	6,801,905
Specification	OLS	OLS	OLS

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. We present the estimations with OLS specification here to have cross-column comparisons in a clearer way, and the estimations with the logit specification reveal a similar pattern. We repeat the analyses on the marketing departments where the employees focus most on the professional questions, and the results are also consistent. The estimation results are omitted here for parsimony. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

The role of job function proximity

We further analyze how having similar job functions is related to the knowledge-sharing probability along with being at the same location. We focus on dyads in which both employees are in marketing departments and classify employees into four categories based on their job functions and work locations. We define three new binary variables: $SameLocSameFunc_{ij}$ equals 1 if provider *i* and seeker *j* have the same job function and work at the same location, and 0 otherwise; $DiffLocSameFunc_{ij}$ equals 1 if provider *i* and seeker *j* have the same job function but work at the different location, and 0 otherwise; $SameLocDiffFunc_{ij}$ equals 1 if provider *i* and seeker *j* work at the same location but have different job function, and 0 otherwise. The three binary variables indicating location proximity and job function help us understand how being at the same work location and having the same job function affect the answer probability. The benchmark describes the scenarios where the provider and seeker have different job functions and work locations.

Table C.7 presents the results for the marketing subsample in Model 1. The significantly positive coefficients of $SameLocSameFunc_{ij}$, $DiffLocSameFunc_{ij}$, and $SameLocDiffFunc_{ij}$ across indicate that these factors are all positively associated with the likelihood of knowledge sharing. The

result of $DiffLocSameFunc_{ij}$ aligns with the job-function-related inclination observed in Table 8, indicating that being the same job function also fosters knowledge-sharing inclination due to shared background and context although they are from different locations. However, the coefficients of $SameLocSameFunc_{ij}$ (t = 3.603, p < 0.001) and $SameLocDiffFunc_{ij}$ (t = 9.982, p < 0.001) are significantly larger than that of $DiffLocSameFunc_{ij}$. This suggests that being in the same location offers additional motivations for employees to share knowledge, while knowledge-focused motivations play a relatively non-salient role motivating the same-location answer inclination.[1] Model 2 extends the analysis to the full sample and yields similar results.

Table C.7. The Role of Job Function Proximity: Same-Location Answer Inclination

	Marketing Subsample	Full sample
	(1)	(2)
Dependent Var.:	*Answer*	*Answer*
SameLocSameFunc	0.350***	0.288***
	(0.015)	(0.014)
DiffLocSameFunc	0.063***	0.058***
	(0.003)	(0.003)
SameLocDiffFunc	0.532***	0.475***
	(0.009)	(0.009)
QLen	0.038***	0.042***
	(0.005)	(0.004)
Exp_Ans	0.708***	0.702***
	(0.004)	(0.004)
Exp_Ques	0.110***	0.126***
	(0.004)	(0.003)
PreviousInt	0.071***	0.063***
	(0.004)	(0.004)
InterestRele	0.403***	0.403***
	(0.012)	(0.011)
OnlineExptSim	0.719***	0.702***
	(0.017)	(0.016)
Provider FE	YES	YES
Seeker FE	YES	YES
Time FE	YES	YES
Pseudo R^2	0.175	0.176
Observations	16,190,623	18,670,135
Specification	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01.

[1] Despite the higher level of similarities between providers and seekers having the same job function and working at the same location (i.e., $SameLocSameFunc_{ij}$ = 1) compared to those working at the same location but with different job function (i.e., $SameLocDiffFunc_{ij}$ = 1), the estimation of the coefficient for $SameLocSameFunc_{ij}$ (i.e., 0.350) is lower than the estimation of the coefficient for $SameLocDiffFunc_{ij}$ (i.e., 0.532). The difference is significant (t = -2.294, p = 0.022), which further suggests that multiple mechanisms are at play. If it is only about social-related motivations play a role, we should observe that the level of answer inclination increases with the similarity and social closeness among the employees; and if it is only about knowledge-focused motivations, we should observe that the level of answer inclination increases with the similarity among the employees because the level of common background increases with the similarity.

Appendix D. Full Regression Tables for Analyses

Because of the page limitation, we only show the regression results for the main variables for Table 6-9, 11-12, and 14 in the manuscript. In this appendix, we show the full regression for these analyses as follows.

Table D.1. Results for Knowledge Seekers with the Same-Name Feature (Full Results for Table 6)

	(1)	(2)	(3)	(4)
Dependent Variable	*Answer*	*Answer*	*Answer*	*Answer*
SN_SameDep	0.142**	0.158***	0.127**	0.146**
	(0.058)	(0.058)	(0.059)	(0.059)
QLen	0.058***	0.057***	0.063***	0.063***
	(0.014)	(0.014)	(0.014)	(0.014)
Exp_Ans	-0.040***	-0.019	-0.039***	-0.018
	(0.015)	(0.015)	(0.015)	(0.015)
Exp_Ques	0.766***	0.771***	0.767***	0.771***
	(0.015)	(0.015)	(0.015)	(0.015)
PreviousInt	0.143***	0.140***	0.156***	0.156***
	(0.012)	(0.012)	(0.012)	(0.012)
InterestRele	0.027*	0.019	-0.004	-0.006
	(0.014)	(0.014)	(0.015)	(0.015)
OnlineExptSim	0.106**	0.132***	0.105**	0.127***
	(0.043)	(0.044)	(0.044)	(0.044)
AdditionalControls	NO	YES	NO	YES
Seeker FE	YES	YES	YES	YES
Provider FE	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
Observations	1,168,416	1,168,416	1,135,373	1,135,373
Pseudo R^2	0.167	0.168	0.166	0.167
Specification	Logit	Logit	Logit	Logit

Notes. 42 users with the same-name feature have posted 1,203 questions in our study period. In this group of analyses, we focus on knowledge seekers with the same-name feature to exclude the alternative explanation that the observed difference in answer probabilities is attributable to the same-name feature. As a robustness check, we estimate the model with the original full sample, and the results are consistent (but omitted here for parsimony). Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table D.2. Subsample of Question Types (Full Results for Table 7)

Question Type	Product	Customer	General
	(1)	(2)	(3)
Dependent Var.:	*Answer*	*Answer*	*Answer*
SameDep	0.00170***	0.00285***	0.00264***
	(0.00022)	(0.00037)	(0.00026)
QLen	0.00226***	-0.00116*	0.00179***
	(0.00023)	(0.00047)	(0.00028)
Exp_Ans	0.01256***	0.01117***	0.01313***
	(0.00011)	(0.00016)	(0.00013)
Exp_Ques	0.00723***	0.00858***	0.00779***
	(0.00023)	(0.00037)	(0.00028)
PreviousInt	0.00365***	0.00281***	0.00268***
	(0.00029)	(0.00051)	(0.00036)
InterestRele	0.01048***	-0.00056	0.00832***
	(0.00047)	(0.00081)	(0.00057)
OnlineExptSim	0.00224***	0.01068***	0.00904***
	(0.00048)	(0.00078)	(0.00059)
Provider FE	YES	YES	YES
Seeker FE	YES	YES	YES
Time FE	YES	YES	YES
Pseudo R^2	0.10827	0.10934	0.14996
Observations	7,199,115	2,328,865	6,801,905
Specification	OLS	OLS	OLS

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. We present the estimations with OLS specification here to have cross-column comparisons in a clearer way, and the estimations with the logit specification reveal a similar pattern. We repeat the analyses only on the marketing departments where the employees focus most on the professional questions, and the results are also consistent. The estimation results are omitted here for parsimony. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table D.3. The Role of Job Function Proximity (Full Results for Table 8)

	Marketing Subsample	Full sample
	(1)	(2)
Dependent Var.:	*Answer*	*Answer*
SameDep	0.070***	0.062***
	(0.004)	(0.004)
DiffDepSameFunc	0.043***	0.040***
	(0.005)	(0.005)
QLen	0.038***	0.042***
	(0.005)	(0.004)
Exp_Ans	0.707***	0.701***
	(0.004)	(0.004)
Exp_Ques	0.116***	0.130***
	(0.003)	(0.003)
PreviousInt	0.072***	0.065***
	(0.004)	(0.004)
InterestRele	0.404***	0.404***
	(0.012)	(0.011)
OnlineExptSim	0.723***	0.704***
	(0.017)	(0.016)
Provider FE	YES	YES
Seeker FE	YES	YES
Time FE	YES	YES
Pseudo R^2	0.174	0.176
Observations	16,190,623	18,670,135
Specification	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01.

Table D.4. Subsample of Professional and Non-Professional Questions (Full Results for Table 9)

Question Type	Professional	Non-Professional
	(1)	(2)
Dependent Variable	*Answer*	*Answer*
SameDep	0.058***	0.052***
	(0.004)	(0.010)
QLen	0.025***	0.106***
	(0.005)	(0.013)
Exp_Ans	0.703***	0.611***
	(0.004)	(0.010)
Exp_Ques	0.129***	0.158***
	(0.004)	(0.009)
PreviousInt	0.065***	0.059***
	(0.004)	(0.010)
InterestRele	0.439***	0.051*
	(0.013)	(0.031)
OnlineExptSim	0.691***	1.84***
	(0.017)	(0.052)
Provider FE	YES	YES
Seeker FE	YES	YES
Time FE	YES	YES
Observations	15,750,515	2,919,620
Pseudo R^2	0.176	0.181
Specification	Logit	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. The result is consistent if we focus on the leisure-related questions when analyzing the non-professional questions. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table D.5. Work vs. Social Knowledge-Sharing Orientations (Full Results for Table 11)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	*Work*	*Work*	*Social*	*Social*	*Orientation_Work*	*Orientation_Work*
SameDep	0.0009***	-	-0.002***	-	0.074***	-
	(0.0002)		(0.0002)		(0.008)	
SameLoc	-	-0.002***	-	0.0004*	-	-0.049***
		(0.0004)		(0.0004)		(0.017)
QLen	-0.009***	-0.009***	0.015***	0.015***	-0.610***	-0.611***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.008)	(0.008)
Exp_Ans	-0.0007***	-0.0007***	0.0001	0.0001	-0.024***	-0.024***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.007)	(0.007)
Exp_Ques	0.0008***	0.0008***	-0.0004*	-0.0004*	0.028***	0.028***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.007)	(0.007)
PreviousInt	-0.002***	-0.002***	0.001***	0.001***	-0.088***	-0.088***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.008)	(0.008)
InterestRele	0.018***	0.018***	-0.020***	-0.020***	0.958***	0.958***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.024)	(0.024)
OnlineExptSim	-0.003***	-0.003***	0.015***	0.015***	-0.469***	-0.467***
	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.023)	(0.023)
Provider FE	YES	YES	YES	YES	YES	YES
Seeker FE	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
R^2/Pseudo R^2	0.046	0.046	0.057	0.057	0.048	0.048
Specification	OLS	OLS	OLS	OLS	Logit	Logit

Notes. The analysis includes 623,084 knowledge-sharing interactions. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table D.6. Interaction of Department and Location Proximity (Full Results for Table 12)

	(1)
Dependent Variable	*Answer*
SameDepSameLoc	0.172
	(0.109)
SameDepDiffLoc	0.064***
	(0.008)
DiffDepSameLoc	0.474***
	(0.043)
QLen	0.042***
	(0.006)
Exp_Ans	0.701***
	(0.023)
Exp_Ques	0.127***
	(0.024)
PreviousInt	0.061***
	(0.010)
InterestRele	0.403***
	(0.025)
OnlineExptSim	0.703***
	(0.077)
Provider FE	YES
Seeker FE	YES
Time FE	YES
Observations	18,670,135
Pseudo R^2	0.175
Specification	Logit

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

Table D.7. Features of Same-Department vs. Non-Same-Department Answers (Full Results for Table 14)

	(1)	(2)	(3)	(4)
Dependent Variable	*NumNonStop*	*AnsQueRelevance*	*BestAns*	*RespTime*
SameDep	0.003*	0.003**	0.069*	0.006**
	(0.001)	(0.001)	(0.035)	(0.002)
Qlen	0.065***	0.0006	-0.061	0.030***
	(0.002)	(0.002)	(0.040)	(0.003)
Exp_Ans	-0.048***	0.0008	-0.250***	-0.161***
	(0.001)	(0.001)	(0.028)	(0.003)
Exp_Ques	-0.008***	-0.0006	-0.014	0.009***
	(0.001)	(0.001)	(0.027)	(0.002)
PreviousInt	2.54e-5	0.003*	0.129***	0.011***
	(0.001)	(0.002)	(0.035)	(0.002)
InterestRele	0.099***	-0.004	0.013	-0.089***
	(0.004)	(0.004)	(0.102)	(0.007)
OnlineExptSim	-0.032***	0.002	-0.123	-0.282***
	(0.006)	(0.003)	(0.132)	(0.011)
Provider FE	YES	YES	YES	YES
Seeker FE	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
Observations	687,083	687,083	687,083	687,083
R^2	0.266	0.019	0.145	0.384
Specification	OLS	OLS	Logit	OLS

Notes. Heteroskedasticity-consistent robust standard errors are in parentheses. *$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

References

Bai, X., Marsden, J. R., Ross, W. T., & Wang, G. (2020). A note on the impact of daily deals on local retailers' online reputation: Mediation effects of the consumer experience. *Information Systems Research*, *31*(4), 1132-1143.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61-84.

Geva, M., Goldberg, Y., & Berant, J. (2019). Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics.

Kuk, G. (2006) Strategic interaction and knowledge sharing in the KDE developer mailing list. *Management Science*, 52(7), 1031-1042.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations*.

Puranam, D., Narayan, V., & Kadiyali, V. (2017). The effect of calorie posting regulation on consumer opinion: A flexible latent Dirichlet allocation model with informative priors. *Marketing Science*, *36*(5), 726-746.

Song, Y., Shi, S., Li, J., & Zhang, H. (2018). Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *2*, 175-180.

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press.

Tian, Q., Zhang, P., & Li, B. (2013). Towards predicting the best answers in community-based question-answering services. In *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 725-728.

Wu, J., Zheng, Z. (Eric), & Zhao, J. L. (2021). FairPlay: Detecting and deterring online customer misbehavior. *Information Systems Research*, *32*(4), 1323-1346.

Yang, F., Yih, W. T., & Meek, C. (2014). Wiki Q&A: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, *2015*, 2013-2018.

Yu, Y., Yang, Y., Huang, J., & Tan, Y. (2023). Unifying algorithmic and theoretical perspectives: Emotions in online reviews and sales. *MIS Quarterly*, *47*(1), 127-160.

Xu, B., Xing, Z., Xia, X., Lo, D., & Li, S. (2016). Domain-specific cross-language relevant question retrieval. *Empirical Software Engineering*, *22*(2), 656-680.