Attention Harvesting for Knowledge Production

by

Fan Yu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved September 2019 by the
Graduate Supervisory Committee:

Marco Janssen, Chair
Carlos Castillo-Chavez
Yun Kang

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

This dissertation seeks to understand and study the process of attention harvesting and knowledge production on typical online Q&A communities. Goals of this study include quantifying the attention harvesting and online knowledge, damping the effect of competition for attention on knowledge production, and examining the diversity of user behaviors on question answering. Project 1 starts with a simplistic discrete time model on a scale-free network and provides the method to measure the attention harvested. Further, project 1 highlights the effect of distractions on harvesting productive attention and in the end concludes which factors are influential and sensitive to the attention harvesting. The main finding is the critical condition to optimize the attention harvesting on the network by reducing network connection. Project 2 extends the scope of the study to quantify the value and quality of knowledge, focusing on the question answering dynamics. This part of research models how attention was distributed under typical answering strategies on a virtual online Q&A community. The final result provides an approach to measure the efficiency of attention transferred into value production and observes the contribution of different scenarios under various computed metrics. Project 3 is an advanced study on the foundation of the virtual question answering community from project 2. With highlights of different user behavioral preferences, algorithm stochastically simulates individual decisions and behavior. Results from sensitivity analysis on different mixtures of user groups gives insight of nonlinear dynamics for the objectives of success. Simulation finding shows reputation rewarding mechanism on Stack Overflow shapes the crowd mixture of behavior to be successful. In addition, project proposed an attention allocation scenario of question answering to improve the success metrics when coupling with a particular selection strategy.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to all those who provided me the possibility to complete this dissertation. A special gratitude I give to Dr Marco Janssen, my advisor, for the help and advising on research and writing. Furthermore I would also like to acknowledge with much appreciation the crucial roles of Dr Carlos Castillo-Chavez and Dr Yun Kang for advising and inspiration on the research. A special thanks goes to Dr Anuj Mubayi, Dr Victor Moreno, Dr Baltazar Espinoza, and Dr Soodeh Boroojeni who helps me in many aspects along the road of PhD degree completion. Last but not least, many thanks go to my peers Dustin Padilla and Juan Renova for valuable suggestion on finding my research interest and my advisor.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

1.1 Crowdsourcing

We have entered a new era with the advent of global internet and the ease of high speed information sharing. The traditional pattern of knowledge propagating and preserving shifts. For examples companies like Digg, YouTube, Stack Exchange and Wikipedia successfully establish. YouTube with millions of users contributes to roughly 37% of all downstream traffic on the mobile internet as of March 2019[1]. Creation and uploading videos is one typical way of crowdsourcing through YouTube but the phenomenon of crowdsourcing can be extended to engage a huge crowd to practice on a common goal of providing answers, ideas and solutions. The reasons of people participating crowdsourcing are getting attention, recognition or financial gain. No matter which intention they possess, the final product of crowdsourcing is a public good which is non-excludable and nonrivalrous(Bade and Parkin, 2007). It characterizes in two ways that this good can be acquired by anyone for free whenever needed and its consumption by others does not affect how much left for others.

1.2 Attention Harvesting

Cognitive surplus refers to the energy and spare time that individuals have above and beyond satisfying the daily necessities of life and can spend on voluntary activities(Shirky, 2010). Democratic surplus is a component of cognitive surplus which is

[1]"YouTube Usage Comprises 37% Of All Mobile Web Traffic, Study Finds", Geoff Weiss, Tubefilter.com, https://www.tubefilter.com/2019/03/27/youtube-37-percent-all-mobile-traffic/

the "effort, goodwill, expertise, innovation and leadership" that individuals possess and can voluntarily exercise within their spare time and energy(Kelley and Johnston, 2012). During crowdsourcing activities, people devote their capacities of cognitive surplus and democratic surplus into production of public good. During our study, we refer the cognitive surplus and democratic surplus spent on online communities of knowledge as attention and online communities of knowledge trying to attract more attention from people as attention harvesting which redeems attention as a renewable resource and can be harvested from people.

1.2.1 Online Communities of Knowledge

Online communities of knowledge as the main focus of our research refer to the platforms of internet-enabled collective intelligence, whereby large crowd from diverse backgrounds disregarding the vast geographic ranges can willingly collect and transform numbers of tiny contributions into meaningful good for the world(Brabham, 2008). Typical examples of online communities are internet encyclopedia Wikipedia, question and answering site Stack Exchange(https://stackexchange.com/) and code sharing platform GitHub(https://github.com/).

We conduct the research and modeling based on Stack Exchange type of online community. This is a pull system of question and answering site that people can search and choose the interested questions to answer and view. Individual is actively "pulling" the interested topics out from the site instead of topics being recommended to users or "pushed" by the system. News aggregator site Digg(www.digg.com) is an example of another type of online community with a push system. In latest 2019 statistics of Stack Exchange (https://stackexchange.com/about), there are 50.7 million unique visitors monthly and total 9.6 billion of pageviews. The total registered users is about 9.5 million. The Stack Exchange network consists of 133 Q&A com-

munities including its flagsite Stack Overflow. Stack Overflow alone serves more than 50 million developers every month.

Stack Overflow's success shall credit to its reputation and privilege system. Users get awarded for reputation score by asking good questions and providing good answers. Awarding process will distribute different reputation scores whenever there is someone voting up or accepting the answer. Higher reputation score will unlock levels of privilege. Users with corresponding privilege can tip the good questions and answers by voting up and voting down for bad answers as reputation punishment. Besides the good awarding mechanism to stimulate contribution and engagement, Stack Overflow also have linked related question, duplicate and similar questions together and formed a question hierarchy so that users can easily see and explore related topics with similar interest.

1.2.2 Other Ways of Attention Harvesting

Other major approaches to harvest people's attention will be advertisement and online competition. Advertisement is very common and widely used to populate information for monetary intention. As matter of fact, it is a low-efficient way to harvest attention. As reported in 2018[2], clickthrough rate across all ad formats and placements display ad is just 0.05%. Furthermore as far as I know the attention harvested by advertisement didn't directly produce public good which differentiate itself from other approaches of harvesting.

For online competitions, multiplayer online game Foldit and Netflix Prize for best movie recommendation are two well-known and successful examples. The online game

[2]"US, Europe and Worldwide display ad clickthrough rates statistics summary", Dave Chaffey https://www.smartinsights.com/internet-advertising/internet-advertising-analytics/display-advertising-clickthrough-rates/

Foldit allows players to collaborate and compete to find out the accurate protein structure. Surprisingly algorithm developed by Foldit players outperforms previously published methods and shows striking similarity to unrevealed and independent work by scientists(Khatib *et al.*, 2011). In order to improve costumers' satisfaction to the recommended movies, Netflix awarded one million grand prize started in 2006 to find the best algorithm of recommendation system based on costumers' preference. Netflix Prize (https://www.netflixprize.com/) motivated fast growing for the field of machine learning. The creation of new recommendation algorithms also benefits to solve other problems.

1.3 Why is Studying the Attention Harvesting and Public Good Production Process so Important?

The emergence of participatory platforms of crowdsourcing brings many potential opportunities in legitimacy, government, information technology and network society sectors to enlighten, engage and empower the public democratic surplus. While many initiatives have established to maximize the opportunities, encourage civic participation and transit from government to governance, underlying problems like regulation on public good being created and resource management of democratic surplus become critical and urgent. As limited and renewable resource, our attention is uneasy to harvest and become democratic surplus after surviving from distractions in daily life and work. We shall spend this surplus wisely and efficiently. Therefore understanding the process of harvesting attention and transmission into production of public good becomes so important to be able to efficiently utilize this civic resource and maximize the social null. The study brings insight of production process and helps policy maker, government and platform organizers more effectively and efficiently direct the crowdsourcing power into the most needed and beneficial place.

1.4 Projects Outline

Dissertation tries to understand and study the process of attention harvesting and knowledge production on typical online Q&A communities. Goals of this study include quantifying the attention harvesting. In project 1, we start with a simplistic discrete time model on a scale-free network and provide the method to measure the attention harvested. Further project 1 highlights the effect of distractions on harvesting productive attention and in the end concludes which factors are influential and sensitive to the attention harvesting. In project 2, we extend the scope to quantify the value and quality of knowledge with focus on the questions answering dynamics. We model how attention was distributed under typical answering strategy on a virtual network of online Q&A community. While questions are being answered, quality and value are computed. The final result provides approach to measure the efficiency of attention transferred into value production and examine the contribution of different scenarios under various computed metrics. For the project 3, it is an advanced study on the foundation of virtual question answering community from project 2. With highlights of modeling different user types, behaviors and preferences, we stochastically simulate individual decision and behavior. Our work investigates the contribution by each user strategies and perform sensitivity analysis on different mixture of user groups.

1.4.1 Project 1:Harvesting Attention to Produce Knowledge under Distractions

Nowadays fragmentation of attention becomes an urgent problem[3]. How would this affect the online knowledge production and crowdsourcing behavior? This study

[3]"Multitasking: Focus and Dispersion in the Age of FOMO", NURUN https://www.nurun.com/en/our-thinking/emerging-behavior/multitasking-focus-and-dispersion-in-the-age-of-fomo/

about question and answer community (QAC) models the dynamics of attention harvested into knowledge production and build on top of the scale-free social network generated from empirical study. Sensitivity analysis is performed to find the most critical and sensitive parameters in the model and network structure shows little effect on final amount of attention harvested. Further simulation result shows there is a conditional optimum by reducing network linkage from question to question. The condition of result is depended on the likelihood of staying at the system when being given less number of attractive choices.

1.4.2 Project 2: Return On Assets of Attention Harvesting for Knowledge Production

We look into the measurement of efficiency for attention harvesting and knowledge production and we also care about the effect of typical answering strategies on overall community performance. The economic concept of Return On Assets(ROA) is adapted to estimate the return of value from devoting attention into answering question. This research constructs the virtual environment of online question and answers community and simulate attention distribution to answer questions under two answering strategy groups identified by previous studies. The preliminary finding shows distinct contribution patterns between two strategy groups on the overall questions answered percentage, average quality and value across the community. In sum, group in favor of difficult questions to answer contributes high quality and high value questions but overshoots the questions with too many answers and thus less value per answer, while group in favor of easy questions to answer helps improve the overall question answered percentage and contributes to the most of mid-level quality and value questions.

1.4.3 Project 3: How Behavior of Users Impacts The Success of Online Q&A Communities

What makes Question & Answer (Q&A) communities productive? In this article we look how diversity of behavioral types of agents impact the performance of Q&A communities using different performance metrics. We do this by developing an agent-based model informed by insights from previous studies on Q&A communities. By analyzing different strategies for how questions are selected and answered we find that there mixtures of strategies leading to best outcomes for different performance conditions. Nevertheless, Q&A communities that reward participants to focus on answering the new questions have the best performance in answering questions, improving quality and solving difficult tasks, which is in line with observed outcome. In conclusion we find that the current strategies of Q&A answers are in line with high performance of producing public benefit from the collective attention available.

Chapter 2

HARVESTING ATTENTION TO PRODUCE KNOWLEDGE UNDER DISTRACTIONS

2.1 Introduction

Attention is becoming an expensive commodity due to intense competition for our attention in low cost ways. A well-connected digital world and the internet appropriate our attention by showing abundant and fast-producing content online. Long and continuous focus on one task becomes extremely rare[1]. Typical online behaviors are either engaging a serial of shifting from task to task or multitasking. Both behaviors characterize the fragmentation of attention which has been studied to be significant on mobile devices under various environmental distractions(Oulasvirta, 2005). Fragmentation of attention is not beneficial to complete tasks and be productive on the online knowledge producing communities like Stack Exchange, Wikipedia and so on. Understanding the dispersion of attention under influence from over-redundant and distracted content becomes critical because crowdsourcing is an emerging trend nowadays. A measurement for the cost of online knowledge producing can help to improve efficiency and better utilize the power of crowdsouring.

This study aims at modeling the attention harvesting on online Q&A communities and puts particular interest in attention dispersion under distraction of multiple linked topics and contents. We investigate the cost of knowledge production and hope

[1]"Multitasking: Focus and Dispersion in the Age of FOMO",NURUN https://www.nurun.com/en/our-thinking/emerging-behavior/multitasking-focus-and-dispersion-in-the-age-of-fomo/

to provide insight of operations efficiency for developers and organizers of online communities. The study on attention dispersion under effect of fragmentation of attention provides insight and motive for research interest in the chapter 4 while we are looking into users answering strategy of attention allocation.

Wang *et al.* (2016); Wu and Janssen (2015) find Q&A communities similar to our topic of interest follow typical scale-free networks. The scale-free network follows the critical rule of power-law distribution of edges and nodes(Barabási and Bonabeau, 2003). Our implementation of generating a scale-free network to study is mixing of preferential attachment and reversed preferential attachment(Wu and Janssen, 2015). Our model is a difference model of discrete time system and time spent at each node as residences' ages are recorded and labeled into discrete stages. With empirical studies(Rutz and Bucklin, 2012; Nair, 2010; Kim *et al.*, 2011; Park, 2017; Huberman *et al.*, 1998; Liu *et al.*, 2010) on distribution of user behaviors through time, we look into final steady status of the network and define an objective function about harvested attention for production to observe.

We perform sensitivity analysis on important model parameters. One of the findings is that the probability of exiting the network is the most sensitive factor to impact the value of pre-defined objective function, which is measuring the weighted sum of population on the network. The mixing relation between preferential attachment and reversed preferential attachment only determines the network structure but has no impact on the objective function of harvested attention. Further we investigate the effect of distraction by optimally removing edges from the network. Finding shows taking out selected edges from a standard genetic algorithm optimization will maximize the value of our objective function until an optimum of objective function is reached after that objective function is decreasing. The optimum of objective function conditions on optimal network connection from genetic algorithm. Such an

optimum would not exist if the likelihood of staying at the original node decreases below a threshold. A parameter called stickiness factor captures the dynamics of this likelihood and show this threshold is close to value of 0.7. When value of stickiness factor is 0.7, people who would have switched to a removed node will now have an alternative probability of staying at current node with the value of 70% of original staying probability. Decisions are made at every 10 seconds in the model.

2.2 Model and Method

We are interested in how much attention is harvested on the network. In our model the assumption is that attention is time spent on the network. We keep tracking the users population and how much time they spend on the network in the model. Let $P_i(t, a)$ be the population of users at node i, time t and residence's age a. Time spent on each node is recorded and labeled into discrete time stages called residence's age. Each topic of interest or question here is referred as node. Linkages between topics and questions are edges on the virtual network. The total number of nodes in the network is N. The time period we study is 500 seconds in total which is relatively short compared to the period of network development and question creation. We consider the problem as a fixed network and not growing. The total number of age bins is K.

$$\begin{cases} P_i(t+1,1) = \lambda_i + s_{ii} r(1-c) P_i(t,1) + \sum_{a=1}^{K} \sum_{j \neq i}^{N} s_{ji} r(1-c) P_j(t,a) & \text{for } a = 1 \\ \\ P_i(t+1,a) = s_{ii} r(1-c) P_i(t,a) + r \cdot c P_i(t,a-1) & \text{for } a > 1 \end{cases}$$

Table 2.1: List of Parameters

Parameter	Description
λ_i	Newly arrived population from out of the network to node i
s_{ij}	Switching probability from node i to j
r	Probability of individual did not leaving the system
c	Aging proportion at residence's age bin a

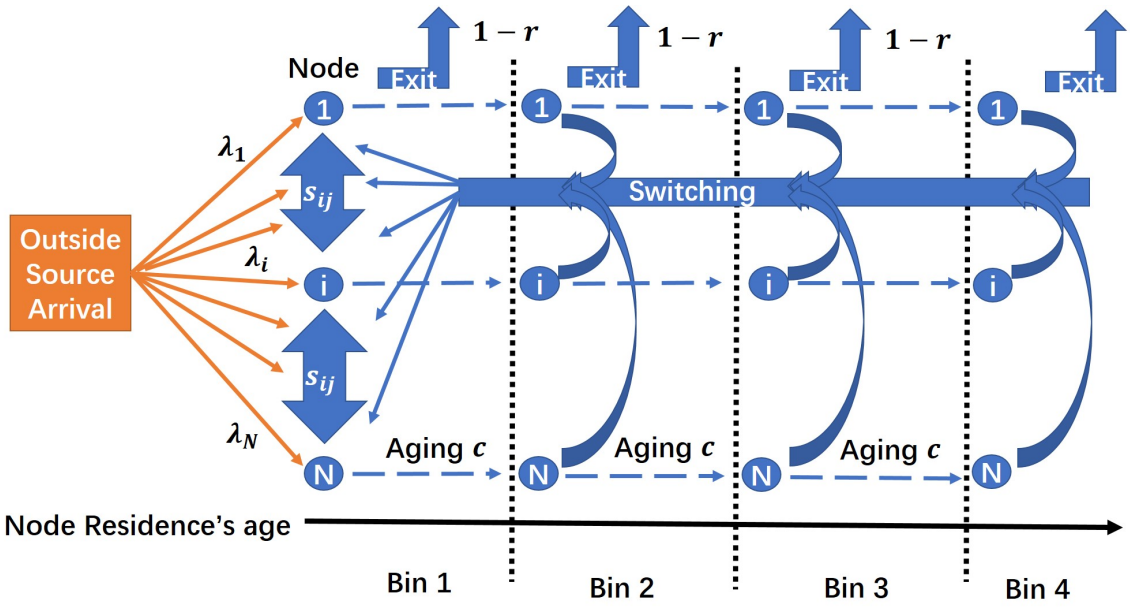

Figure 2.1: Model Flowchart

$P_i(t, 1)$ is population at node i and residence's age bin 1 and progressing in sequence to residence's age bins 2, 3, 4 and so on. Aging population at each time step t is a proportion c of whole population. Aging proportion parameter c depends on the lengths of the age bins L that $c = 1/L$. In out study the length $L = 10$ so that $c = 0.1$. Equally, we group all the actions within a 10-sec bin. As shown in the

model flowchart, when users switch a topic of interest and move to a new node, the corresponding residence's age is reset and start counting from the beginning at bin 1.

Let $\vec{P^*}(a)$ be the column vector of the final population at steady state for all the nodes at residence's age bin a. Row element of vector $\vec{P^*}(a)$ is the final steady population at each node. Let S be the $N \times N$ matrix of all the transmission probability s_{ij} that is the probability from node i to j. Thus the row sum of matrix S is $\sum_j^N s_{ij} = 1$. Note that the diagonal of matrix $S(s_{ii} = 0)$ are all zeros meaning no switching between i to i itself. Finally the vector $\vec{\lambda}$ is the vector of newly arrived population at all nodes. Vector $\vec{\lambda}$ can be seen as the users outside the network being directed to the nodes.

$$\begin{cases} P(t+\vec{1}, 1) = \vec{\lambda} + Diag(S)r(1-c)P(\vec{t}, 1) \\ + \sum_{a=1}^{K} S^T(s_{ii} = 0)r(1-c)P(\vec{t}, a) & \text{for } a = 1 \\ \\ P(t+\vec{1}, a) = Diag(S)r(1-c)P(\vec{t}, a) + r \cdot cP(t, \vec{a-1}) & \text{for } a > 1 \end{cases}$$

It is easy to see such a steady state exist if any probability of leaving the system $r_i(a) > 0$ so that there is a proportion of population leaving the system. The population newly arriving into the network is $\vec{\lambda}$ and is constant. The network will reach the steady equilibrium if the product of population and proportions of population leaving the system equals the newly coming population $\vec{\lambda}$ from outside.

Let such a non negative steady state be $\vec{P^*}(a)$ for $a = 1, 2, 3, ..., K$. We shall have:

$$\begin{cases} \vec{P^*}(1) = \vec{\lambda} + Diag(S)r(1-c)\vec{P^*}(1) + \sum_{a=1}^{K} S^T(s_{ii} = 0)r(1-c)\vec{P^*}(a) & \text{for } a = 1 \\ \\ \vec{P^*}(a) = Diag(S)r(1-c)\vec{P^*}(a) + r \cdot c\vec{P^*}(\vec{a-1}) & \text{for } a > 1 \end{cases}$$

Solve that

$$\begin{aligned} \vec{P^*}(a) &= [I - Diag(S)r(1-c)]^{-1} r \cdot c\vec{P^*}(a-1) \\ &= \prod_{k=2}^{a} \{[I - Diag(S)r(1-c)]^{-1} r \cdot c\} \vec{P^*}(1) \text{ for } a > 1 \end{aligned}$$

In sum,

$$\vec{P^*}(a) = f^{(a)} \vec{P^*}(1) \tag{2.1}$$

$$\begin{cases} f^{(1)} = 1 & \text{for } a = 1 \\ f^{(a)} = \prod_{k=2}^{a} \{[I - Diag(S)r(1-c)]^{-1} r \cdot c\} & \text{for } a > 1 \end{cases}$$

Let $\Phi = \sum_{a=1}^{K} S^T(s_{ii} = 0) r(1-c) f^{(a)}$

$$\vec{P^*}(1) = \vec{\lambda} + Diag(S)r(1-c)\vec{P^*}(1) + \Phi \vec{P^*}(1)$$

Solve that

$$\vec{P^*}(1) = [I - Diag(S)r(1-c) - \Phi]^{-1} \vec{\lambda} \tag{2.2}$$

From the analytical solution of difference model, we find:

- Long term steady population $\vec{P^*}(a)$ for $a > 1$ only depends on lone term steady population at the first residences' age bin $\vec{P^*}(1)$, the aging proportion c, the probability of exiting the system r and only diagonal elements of switching matrix S.

- There is a fixed decaying factor $[I - Diag(S)r(1-c)]^{-1} r \cdot c$ from previous residences' age bin $a-1$ to current residences' age bin a $(a > 1)$.

- Long term steady population at the first bin $\vec{P^*}(1)$ linearly depends on newly arrived population $\vec{\lambda}$

The simulation results of final population at each residence's age bin and for every node were presented in the section of sale-free network generation 2.2.1.

2.2.1 Harvesting Attention Function

We define the objective function Y, which represents the harvested attention used to solve online problems or answer questions. It is a weighed summation of population at each node across all age bins. Quantitatively objective function Y multiplying length of age bin L is the total productive attention on the network we are interested, but qualitatively because length of bin L is constant we focus on weighted population Y on the network.

$$Y = \sum_{a=1}^{K} \omega_a \sum_{j=1}^{N} P_j^*(a) \tag{2.3}$$

ω_a is the weight for harvested utility from each age bin. The weights are computed by a logistic function which gives 0 when age a is small and 1 when age a is approaching the end. The weight value from logistic function is determined by mean parameter A, steepness parameter b, and shape parameter v which are discussed in detail in the section 2.4.

$$\omega_a = \frac{1}{(1 + A * e^{-b*a})^v} \tag{2.4}$$

The idea of weight ω_a takes into account of increasing utility of attention into solving and value producing when people spend longer attention and time. While more time is spent on the topic of interest, or the node, less people will stay on the node because of lack of interest. Whoever stays longer is contributing more effectively into knowledge production. Such dynamics are also explained in the figure 2.2.

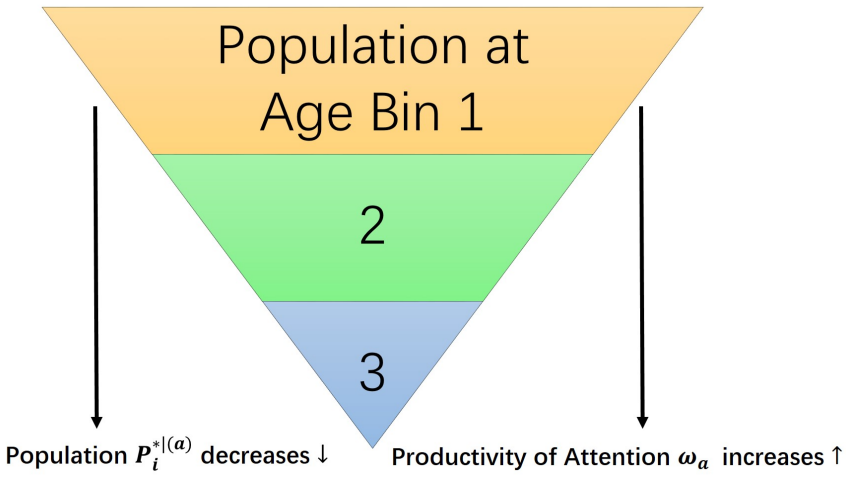

Figure 2.2: Objective Function Diagram

Scale-free Network Construction

The World Wide Web and social networks are scale-free networks and follow characteristic power law for distribution of edges, nodes, and degrees(Albert *et al.*, 2000). Particularly, Wu and Janssen (2015) suggests the mechanism of forming the network of online communities. Furthermore, Stack Exchange is a mixing between preferential attachment and reversed preferential attachment. The concept of preferential attachment is that the attractiveness of an existing node is positively related to its degree. Under reversed preferential attachment, the attractiveness would be inversely proportional to its degree. The best fitted mixing ratio between these two attachment approaches on hundreds of Stack Exchange networks is 0.3 with 30% preferential attachment and 70% reversed preferential attachment (Wu and Janssen, 2015). In this section, we aim at constructing a scale-free network of mixing preferential attachment and reversed preferential attachment based on empirical study and results Wu and Janssen (2015) from over 110 different communities of Stack Exchange.

We first generate a finite network with N=100 nodes and then randomly assign

degree to the initial nodes under power law distribution. Based on the assigned degree, we further determine the number of connectivity for each node. As shown in figure 2.3 and 2.4, the number of nodes and the connectivity of node follows power law of the degree.

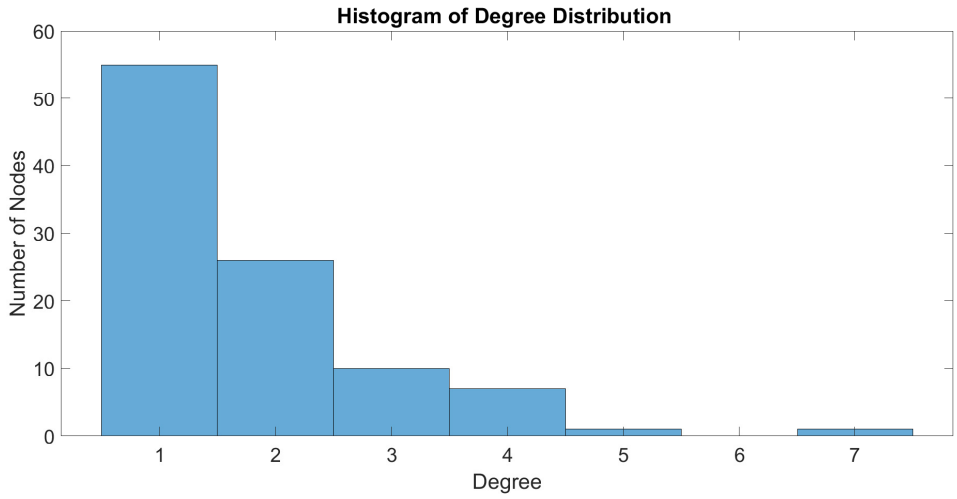

Figure 2.3: Degrees Distribution with N=100 Nodes

Given the best fitted ratio of mixing p=0.3 (Wu and Janssen, 2015), we use the mixing scenario to determine the probability of a node is chosen to be connected and the probability of a node at which users from outside the network arrive:

$$P(i) = p \times \frac{k_i}{\sum_j^N k_j} + (1-p) \times \frac{\frac{1}{k_i}}{\sum_j^N (\frac{1}{k_j})} \tag{2.5}$$

Where $P(i)$ is the probability of choosing node i based on the degree of node k_i.

At the node i, the number of connectivity KNN_i is generated by power law. KNN_i also represents the number of edges coming out from node i. We randomly determine where the edge is connecting to by the probability $P(j)$ where $j \neq i$.

After connecting to the selected nodes, the switching out probability from node i to node j(named as s_{ij}) among the connected nodes is standardized $P(j)$ between

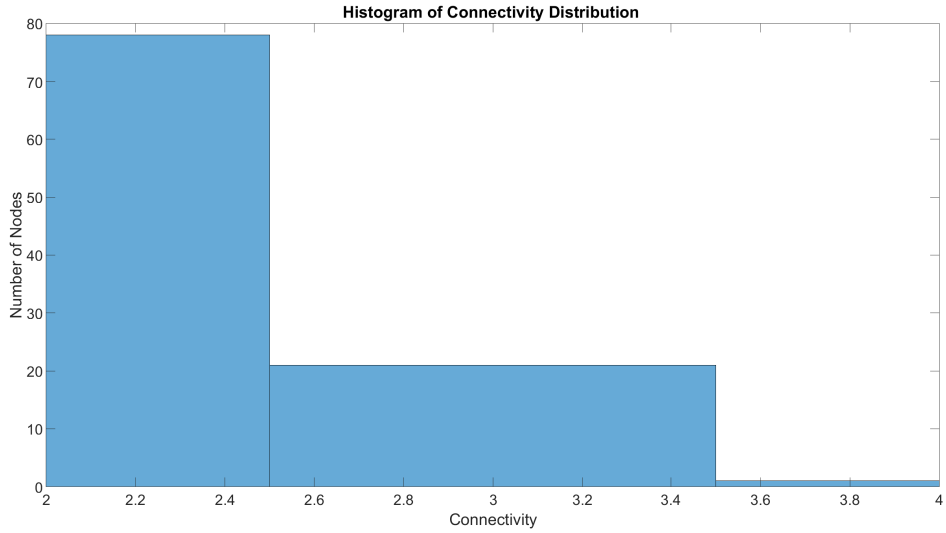

Figure 2.4: Connectivity Distribution with N=100 Nodes

connected nodes. Thus $s_{ij} = 0$ if node i is not connected to node j. $s_{ij} = \frac{P(j)}{\sum_{connectedl}^{N} P(l)}$ when node i connected to node j and l is every connected node to the node i.

In the figure 2.5 below, it is the result of final steady state population at each residence's age bin decaying along the increasing age a. At each age bin, the population is the total from all nodes.

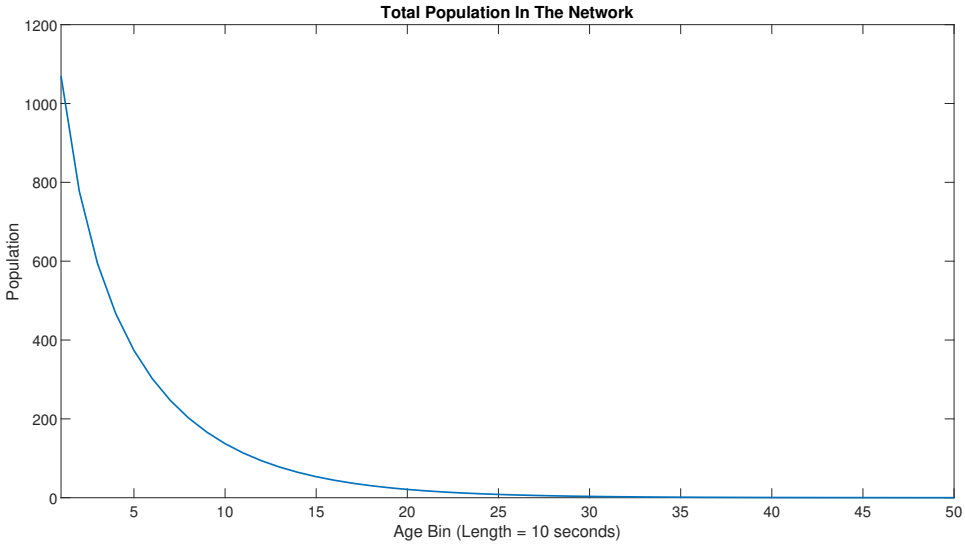

Figure 2.5: Final Population Distribution across Residence's Age Bins

Below in the figure 2.6, it is the histogram of final steady state population at each node with population at each node as x axis and number of node with corresponding population as y axis. As seen in the plot, 15 out of 100 nodes have populations near 50, the highest density point for population on a node. The highest number of final population is close to 90 and lowest at 26.

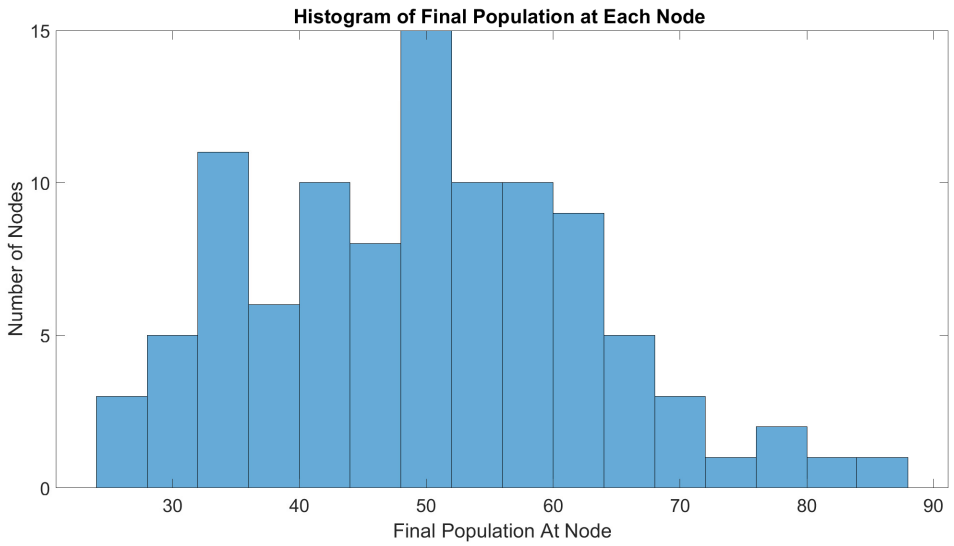

Figure 2.6: Histogram of Final Population at Each Node

2.2.2 Model Parameters

In this section, we discuss the methods of determining the parameter values. As shown below at table 2.2, it lists all the parameters used to build the scale-free network and model. In the Ref column, if it is stated as "Assumed", it is arbitrarily defined by us. We will perform sensitivity analysis to explore the changing effect of all assumed parameters in the model on the objective function Y of harvested attention. There are a few parameters that we reference from previous literature(see table 2.2). The rest were calculated. The section of calculation procedure is also cited in the Ref column.

Table 2.2: Model Parameters

Parameter	Value	Sensitivity Analysis	Ref
λ Newly arrived population from outside	100	Yes	Assumed
s_{ij} Switching probability	0 - 1	No	Calculated2.2.1
r Probability of not exiting	0.98	Yes	Assumed
c Aging factor	0.1	No	Calculated 2.2
$P_L(a)$ Probability of leaving node	0 - 1	No	Calculated2.7
$\lambda_{weibull}$ Scale parameter of Weibull Distribution	20	Yes	Liu *et al.* (2010)
$k_{weibull}$ Shape parameter of Weibull Distribution	0.8	Yes	Liu *et al.* (2010)
p Preferential mixing ratio	0.3	Yes	Wu and Janssen (2015)
ω_a Weight of harvesting	0 - 1	No	Calculated
A Logistic mean parameter	300	Yes	Assumed
b Logistic steepness parameter	0.3	Yes	Assumed
v Logistic shape parameter	2	Yes	Assumed

Note arrived population from outside the network at each node i is $\lambda_i = \lambda \times \frac{P(i)}{\sum_i^N P(i)}$. The proportion $\frac{P(i)}{\sum_i^N P(i)}$ is determined by property of the generated network. The probability of users not exiting the network r is constant across age a and $r = 0.98$. The probability of exiting the network at each residences' age bin is $1 - r = 2\%$.

Liu *et al.* (2010) found that dwelling time on 98.5% of webpages has negative aging effect with shape parameter $k_{weibull} < 1$. Shape parameter controls the probability increasing by time if $k_{weibull} > 1$ or decreasing for $k_{weibull} < 1$. The scale parameter $\lambda_{weibull}$ mostly falls within 400 seconds with 80 percentile at $\lambda_{weibull} = 70$ seconds. Scale parameter controls the mean of the Weibull distribution. In our study, we set two parameter values($\lambda_{weibull} = 20, k_{weibull} = 0.8$) to configure the model. Note that because we group every 10 seconds into one age bin, $\lambda_{weibull} = 20$ in our study equivalent to $\lambda_{weibull} = 200$ in Liu's research(Liu *et al.*, 2010). The probability of leaving a node at age bin a consists of probability of switching nodes from dwelling Weibull distribution and probability of natural exiting out of the system. The difference of leaving probability $P_L(a)$ and exiting probability $1 - r$ is the switching node

probability $Prob_{switch}(a)$ and it is decaying by age.

$$Prob_{switch}(a) = \frac{k_{weibull}}{\lambda_{weibull}}(\frac{a*L}{\lambda_{weibull}})^{k_{weibull}-1}exp(-(\frac{a*L}{\lambda_{weibull}})^{k_{weibull}}) \tag{2.6}$$

$$P_L(a) = 1 - r + Prob_{switch}(a) \tag{2.7}$$

See figure 2.7 for the Weibull distribution curve of leaving probability $P_L(a)$ and constant exiting probability $1 - r$. Function $P_L(a)$ shows the same negative aging curve in article[2].

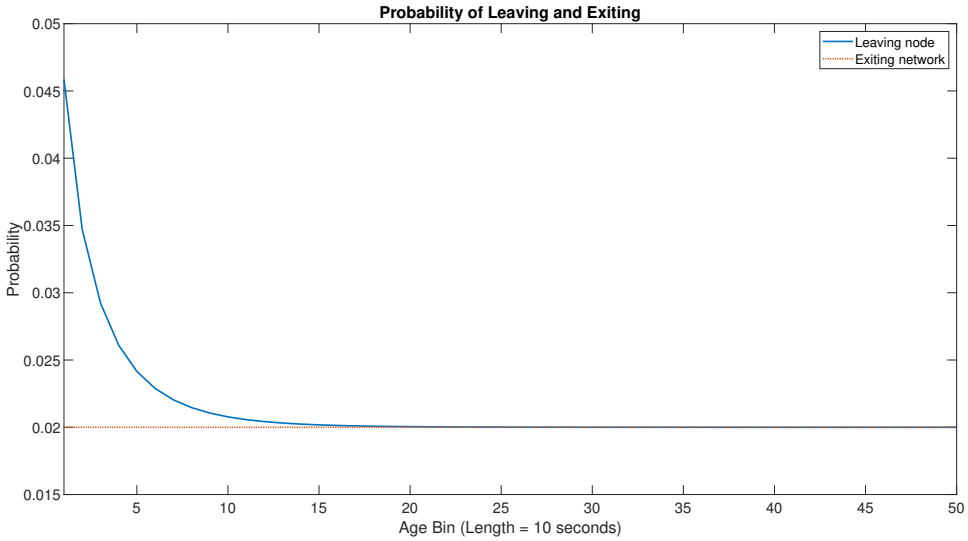

Figure 2.7: Probability of Leaving the Node and Exiting the Network with Parameters $\lambda_{weibull} = 20, k_{weibull} = 0.8$ and $r = 0.98$

The harvesting weight ω_a shall have two characteristics that when age a is close to zero, weight ω_a is close to zero and when age is increasing to one along age a, ω_a is getting larger. We use form of logistic function shown in equation 2.4 to generate the weights in the figure 2.8 .

[2]"How Long Do Users Stay on Web Pages?", Jakob Nielsen https://www.nngroup.com/articles/how-long-do-users-stay-on-web-pages/

A is the mean parameter which controls the sigmoid's midpoint. b is the logistic growth rate parameter which controls the steepness of the curve. v is the shape parameter which decides how fast the curve is approaching to one.

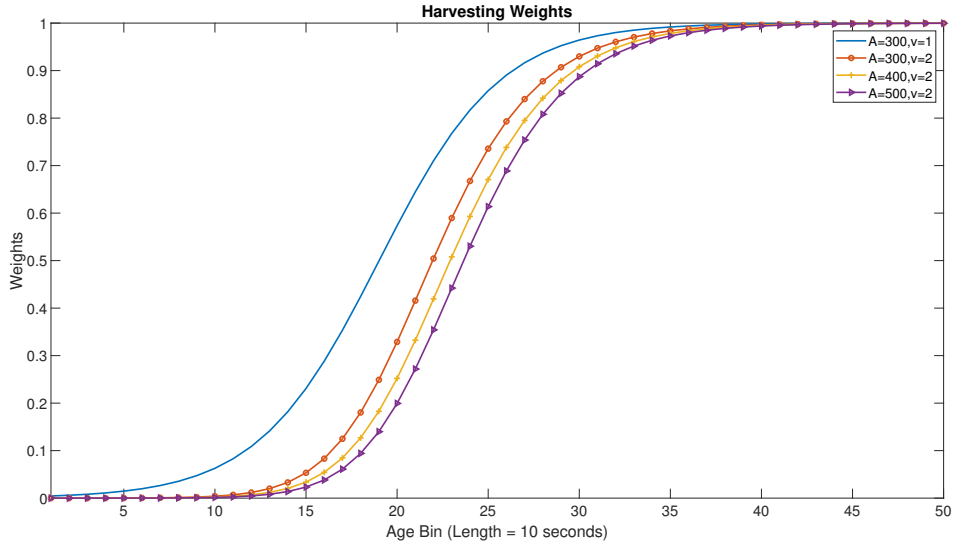

Figure 2.8: Harvesting Weights under Different Parameters and $b = 0.3$

2.3 Result

We first examine the result of sensitivity analysis. During the section 2.3.1, we perform local one-at-a-time sensitivity analysis on eight assumed model parameters. In the next subsection, we are interested in how the population dynamics in the network changes while we reduce edges to switch out in the network.

2.3.1 Sensitivity Analysis

Given the objective function of Y in section 2.2.1, we look at the change of eight model parameters in term of the objective function Y. The sensitivity coefficient is defined as $f = \frac{\Delta Y/Y}{\Delta x/x}$ where f is the sensitivity coefficient and x is the model parameter. The change of parameter and objective value are Δx and ΔY.

Table 2.3: Parameters Table for Sensitivity Analysis

Parameter	Baseline Value	Range	Sensitivity Coefficient	R^2
λ Newly arrived population from outside	100	(50,500)	1	1
r Probability of not exiting	0.98	(0.95,0.999)	1201.4	0.344
$\lambda_{weibull}$ Scale parameter	20	(10,30)	-0.1654	0.989
$k_{weibull}$ Shape parameter	0.8	(0.5,0.99)	0.764	0.954
p Preferential mixing ratio	0.3	(0.01,0.6)	-7.72×10^{-17}	0.0429
A Logistic mean parameter	300	(100,500)	-0.7878	0.897
b Logistic steepness parameter	0.3	(0.1,0.5)	3.1976	0.933
v Logistic shape parameter	2	(1,3)	-1.4017	0.857

In the table 2.3, we see probability of not exiting the network r is the most sensitive parameter with coefficient value 1157 which means that increasing 1% of r will cause 1201.4% increase on Y locally near the neighborhood of baseline parameter set. Especially we can see from the figure of local sensitivity plot 2.9 that parameter r has extreme nonlinear effect near the upper threshold when r increases close to 1. The non-linearity is caused by the population explosion when $r = 1$ meaning no one will leave the network and the network will consistently grow. Both figures 2.9 and 2.10 provide the fitted linear curves to the simulated sensitivity coefficients of all parameters studied in the sensitivity analysis with 95% confidence level.

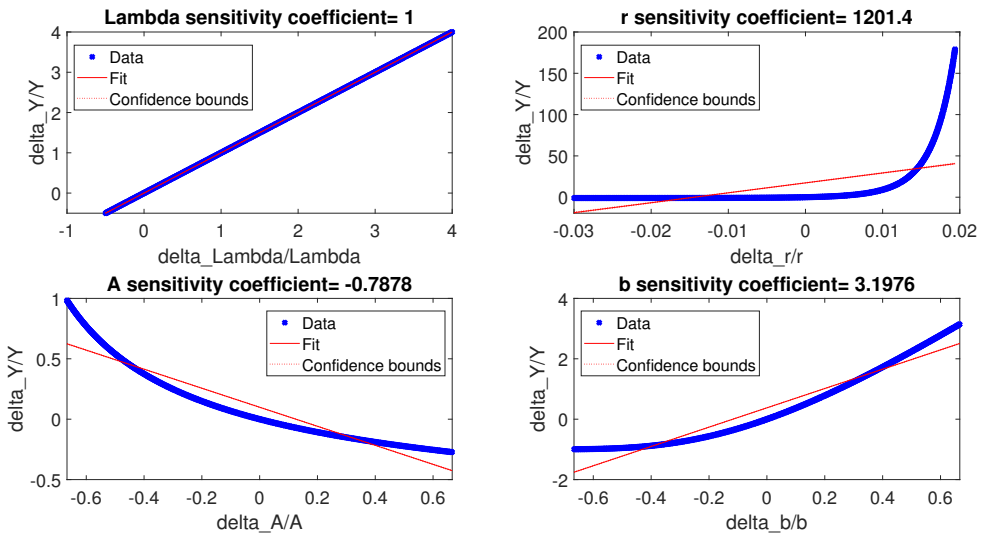

Figure 2.9: Linear Fitted Sensitivity Coefficients for First Four Parameters

The arrived population parameter λ shows perfectly linear trend in figure 2.9 with $R^2 = 1$ and coefficient 1. Increasing 1% of *lambda* will result in 1% increment of Y. Scale parameter of Weibull distribution $\lambda_{weibull}$ shows negative effect to Y because when $\lambda_{weibull}$ goes up, so as the mean of Weibull distribution and more likelihood of switching in the later age. Compared less likelihood of switching in the early age but more likelihood of switching in the later age, less portion of people will stay in long and continuous focus to contribute into objective function Y when $\lambda_{weibull}$ goes up. Similarly for shape parameter of Weibull distribution $k_{weibull}$, the larger the $k_{weibull}$ is, the larger the steepness of the probability density curve of switching probability therefore more likely the switching happens in the early age and less in the later.

Looking into three parameters controlling the weight distribution, we find logistic function steepness parameter b has sensitivity parameter value of 3.1976. Increasing b will lead the weight curve rise up earlier and gives higher weight to the early age. Both of logistic mean parameter A and shape parameter v have negative sensitivity coefficients because increasing the values of both will move the density curve of weight

toward the right side. It means the weight for the early age will decrease and increase for the later age. Considering the population of early age is the majority population and the population of later age becomes minority, both parameters give negative influences on final attention harvesting function Y.

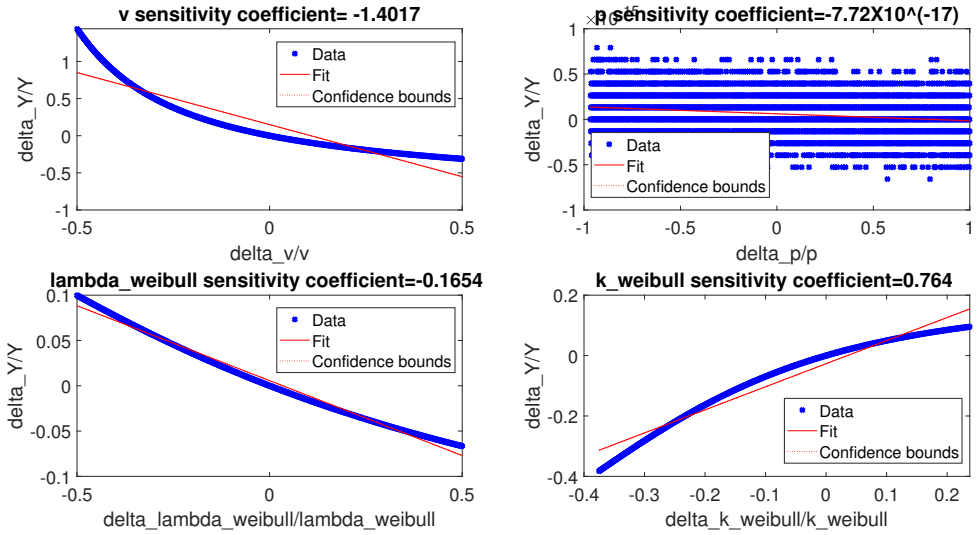

Figure 2.10: Linear Fitted Sensitivity Coefficients for Last Four Parameters

Finally, preferential mixing ratio p has a sensitivity coefficient value close to zero and has non-linearity. It demonstrates p didn't affect the value of attention harvesting function Y significantly. Although preferential mixing ratio is critical parameter to generate a scale-free network and to determine distribution of the newly arrived population to each node, it does not impact the number of users exiting and distribution of population factors across the age $f^{(a)}$. Therefore the total population in the network by age has no relationship with parameter p. It is only important to change population traffic dynamic between nodes.

2.3.2 Optimization of Edge Removal

In this section, we investigate the effect of distraction on attention spent at a node. The research question we would like to answer is if by providing less distracted options we improve productive attention being harvested on the network.

Firstly we assume when people facing less options to switch, they maintain the same logic to make the rest of decision: staying, leaving the network or switching to other nodes. We call this logic behind decisions as proportional distribution. Example is given in the figure 2.11.

Given a connected network, a node is linked to other nodes. There is no transportation to other node when there is no edge linked to others. The likelihoods of users switching from node to node are different and governed by transmission matrix S_{ij}. We study the harvested attention Y and search for critical point if it exists which will maximize the productive attention Y by removing edges other nodes. When removing an edge from m to n at a node m, we assign the probability $s_{mn} = 0$ to the removed edge and re-standardize the transmission matrix S_{ij} so that row sum $\sum_j^N s_{ij}$ equals to 1. The portion $s_{mn}r(1-c)$ individuals who would choose and transmit to node n from node m will now have another chance to make a decision about switching nodes, staying or exiting the network. Then under the assumption that the probability of switching to removed nodes would be proportionally distributed between switching, staying and exiting. In figure 2.11 it is an example of removing one edge from node 3 to node 1. The corresponding probability of switching to node 3 was proportionally distributed into events of switching to other nodes, staying at current node and exiting the network.

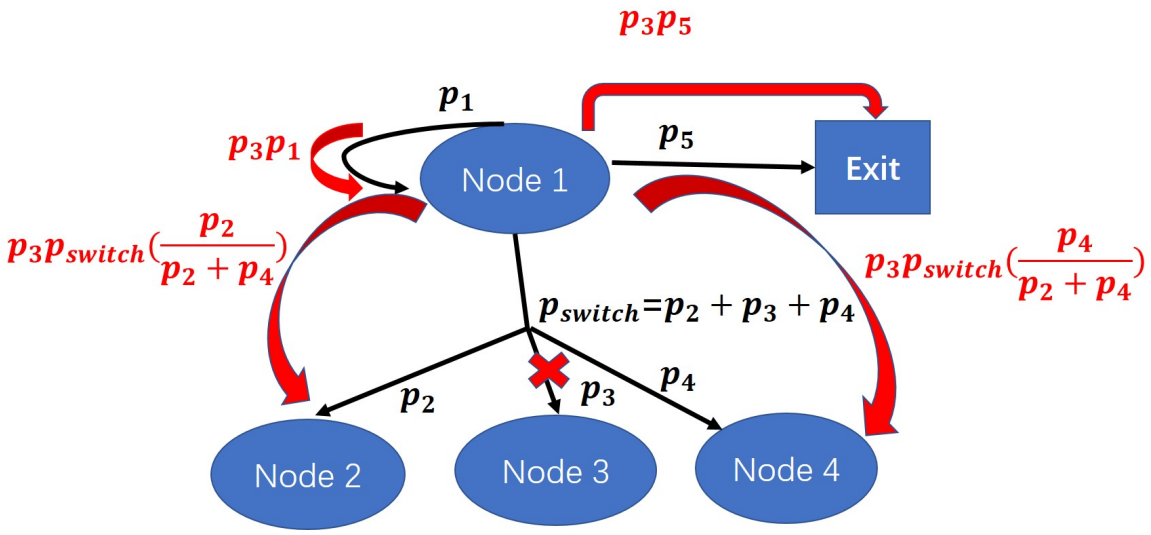

Figure 2.11: Conceptual Chart of Proportional Distribution When Removing a Node: Black Arrows Are the Original Transmitting Probability. Red Arrows Are Additionally the Probability from Removed Node 3 Added into Other Possible Directions. $p_1 + p_2 + p_3 + p_4 + p_5 = 1$.

By reducing the outflow degree of a node, we tend to model the users' decisions under less options and less distraction. We observe the final value of objective function Y to see the change. The method of selecting which edge to remove is a standard genetic algorithm with the goal of maximizing the the objective function Y by searching across the network for the optimal edge. The standard genetic algorithm is implemented in Matlab with built-in GA package and the code of model can be download publicly[3].

Figure 2.11 explain one concept of proportional distribution for the probability of stitching to the removed node. Secondly we explain another concept of disproportional distribution. The idea is using an artifact parameter called stickiness factor to

[3]Matlab source code can be downloaded on CoMSES: https://www.comses.net/codebase-release/dec749e7-3e8c-4dd2-a307-2d47db29297e/

examine the desire to stay at current node when giving less options. Based on the original proportion of switching, exiting and staying. Figure 2.12 gives an example. We simulate and obverse the system under different values of stickiness factor shown in the figure 2.13.

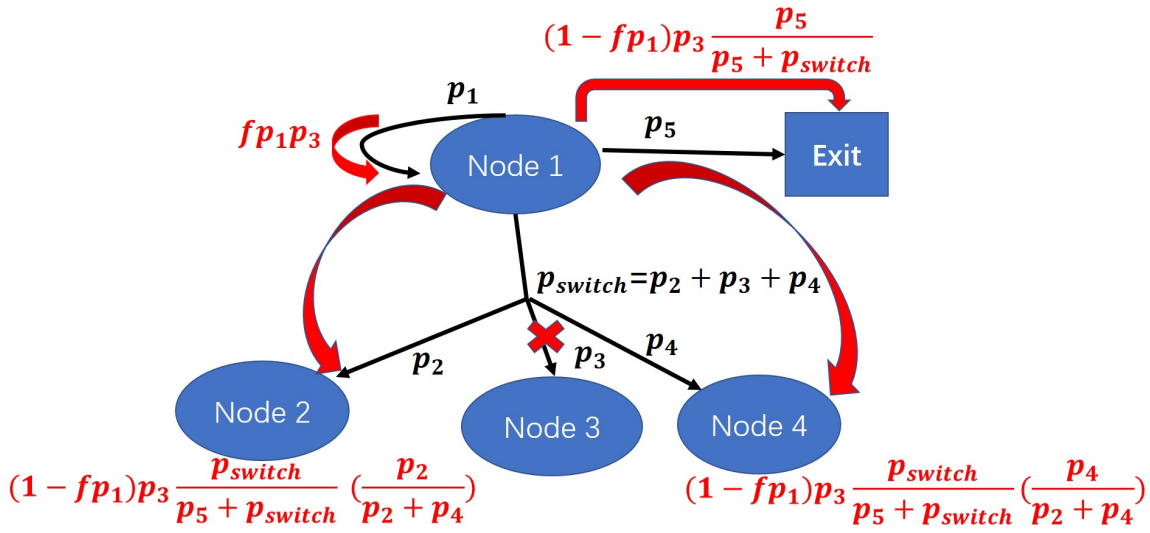

Figure 2.12: Conceptual Chart of Preferential Distribution When Removing a Node: Black Arrows Are the Original Transmitting Probability. Red Arrows Are Additionally the Probability from Removed Node 3 Added into Other Possible Directions. $p_1 + p_2 + p_3 + p_4 + p_5 = 1$. f is the Stickiness Factor Ranged from 0 to 1.

When stickiness factor f equals to 1, it is the same case of proportional distribution. While f decreases, the people would have chosen to switch to the removed node would now instead less likely to stay at current node but more likely to leave the network or switch to other possible nodes. While we are perturbing the stickiness factor f from 0.5 to 1, we see the maximum of objective function Y occurring at different fraction of edges left in the network. Further decreasing stickiness factor f to 0 will result in monotonic decreasing trend for objective function Y.

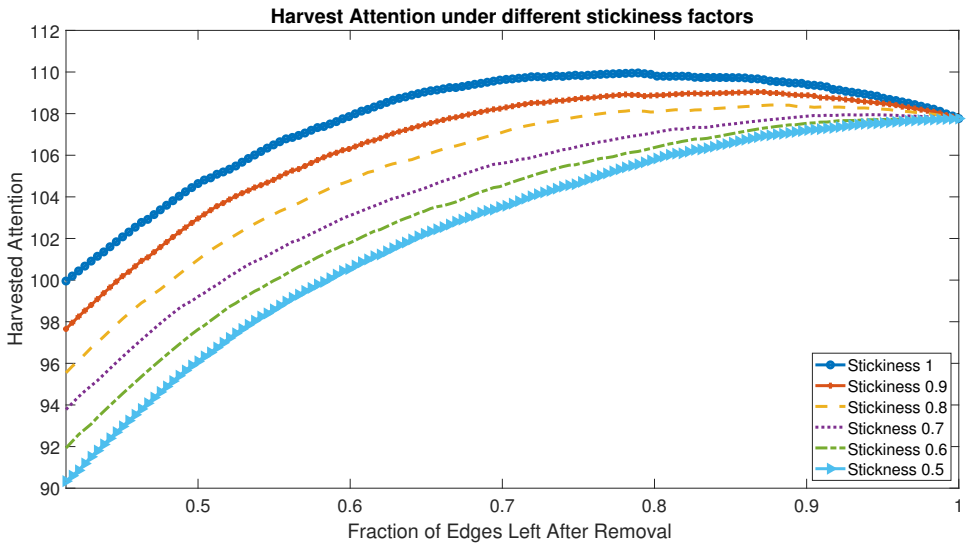

Figure 2.13: Harvested Attention under Different Stickiness Factor from 0.5 to 1

In the figure 2.13, we present the harvested attention curve of objective function Y for different stickiness factor f from 0.5 to 1 under edge removal. When $f = 1$ it is the case of proportional distribution for edge removal. The four lines of $f = 1$,0.9, 0.8, and 0.7 show the objective function Y increasing at first and reaching the maximum before decreasing when fraction of edges left after removal is decreasing. If the likelihood of staying at the original node declines as f decreasing below 0.7, there would be no increasing trend at the beginning of edge removal and the objective function Y is always going down. By taking out the edges, our finding is that, based on the value of stickiness factor f the network has different potential to maximize the objective function Y.

In conclusion, when scale-free network has a high stickiness factor $f \geq 0.7$, by reducing the edges to certain degree we can maximize the harvested attention for production Y. However if the network has low stickiness factor $f < 0.7$, there is no gain to reduce the edges in the network. The optimal structure of the network is depended on the stickiness factor f. As shown in the figure 2.13, the number of edges

to remove before reaching maximum is positively related to the value of stickiness factor f.

2.4 Discussion

The paper addresses the issue of competition for attention and overwhelming attractive contents damping the concentration for knowledge production. Online community tends to provide variety of interesting choices to users in order to attract visiting to the site. However, too much competition for attention results in fragmentation of attention and prevents consistent focus on solving difficult problems. Also it discourages quality sequential computing happening. The work is replied on empirical studies about network and user behaviors to quantify the attention being harvested to produce answers and solution. We perform sensitivity analysis on model parameters and find that the network structure has little impact on the overall quantity of attention harvesting. Also the most sensitive parameter to the final quantity of attention harvesting is the probability of exiting the network which has positive local sensitivity coefficient of 1201.4.

We further investigate the dynamics of reducing distractions on the network by removing edges between nodes. If users are more likely to leave the node when they are given less options, the overall quantity of harvested attention will decline along the edges removal on the network. If users remain the same way of thinking to make the rest of decisions after edge removal, we see the harvested attention actually increases before reaching the maximum and decreases afterward. The maximum of attention is harvested, under the setting that stickiness factor f equals 1, when 20% of the selected edges are removed from the original network. We use the stickiness factor f to model the likelihood of staying at the original node after users are given less choices. The threshold is $f = 0.7$ when harvested attention on the network would not

benefit from edge removal. The finding brings insight and potential to optimize the online community in order to encourage longer concentration and more attention to produce knowledge.

This study has not yet considered various dimensions, complexity of the network and user profile. In the future work, we would like to explore the difference in users expertise, skill set, answering habits and strategies of pursuing reputation scores. As we known, the heterogeneity of user attributes brings different dynamics into the community and users contributes in very different ways. During the optimization of the network to improve the quantity of attention harvesting, we have not yet studied the characteristics of the selected and removed edges fro genetics algorithm. Given a network, the condition of edges to remove remains as an open question. We would like to investigate the connectivity of nodes with removed edges and explore the common pattern of the optimal network structures if more time is permitted.

Chapter 3

RETURN ON ASSETS OF ATTENTION HARVESTING FOR KNOWLEDGE PRODUCTION

3.1 Introduction

There is a 72-hour video being uploaded to YouTube every minute[1]. We are facing exploding amounts of content created while we are hardly able to digest it all. Human attention is renewable but limited resource. It is critical that online users manage their attention spent online in an efficient way. We study attention being used to create knowledge on the online communities among many ways of online attention harvesting. This process is also as known as crowdsourcing. People produce online public good on the variety of topics and interests. In return they look for peer-recognition, self satisfaction, prestige, others' attention or other sources of benefit. In out study, we focus on the crowdsourcing activity on the online Q&A communities. Additionally, the model and methodology potentially can be applied on other online behavioral problems. Knowing the value of those online knowledge created by people is very insightful and helpful.

The goal of this study is to build a standard measurement for the efficiency of online knowledge production on the online Q&A community. Our model helps to provide insight of the gain and lost during attention harvesting into knowledge production. We first summarize previous works on defining and quantifying conceptual

[1]"Multitasking: Focus and Dispersion in the Age of FOMO", NURUN https://www.nurun.com/en/our-thinking/emerging-behavior/multitasking-focus-and-dispersion-in-the-age-of-fomo/

terms related to online knowledge, specifically on the online Q&A communities. On top of the defined metrics, we simulate attention spent on answering questions under two typical online user behaviors on a virtual environment. We compute metrics like quality of question and answers, value of question and answers, and percentage of question answered.

Many studies contributed to quantify online knowledge (Huberman *et al.*, 2009; Anderson *et al.*, 2012; Huna *et al.*, 2016; Baltadzhieva and Chrupała, 2015; Kavaler and Filkov, 2018; Li *et al.*, 2012; Ravi *et al.*, 2014; Liu *et al.*, 2013; Sun *et al.*, 2018). Estimating the value of online knowledge production is a challenging task. Anderson *et al.* (2012) used data mining approaches on Stack overflow data to find the determinants of the value of answers. Number of pageviews, in their study, is the measurement for value of answers to a question. Pageviews is a good measurement because it reflects the demand for the answer to the question and times of consumption. We develop our form of value reflecting this popularity. Given observable variables like the number of answers toward the question, total score of all answers and how long the first answer arrives, Anderson *et al.* (2012) finds significant predicting power to the number of pageviews a year later.

Several papers have addressed the quality of questions and found the determinants of the questions' quality(Li *et al.*, 2012; Baltadzhieva and Chrupała, 2015). For example, Li *et al.* (2012) investigated the entertainment & music category of Yahoo! Answers and gives distribution of the question quality in Movies and Musics sector. In term of the quality of question and answers together, high quality question will more likely to attract high quality answers and more attempts to answer(Baltadzhieva and Chrupała, 2015). Difficult questions get more quality answers as well based on the behavioral study of answering (Wu *et al.*, 2016; Yang *et al.*, 2014). One popular method of evaluating the answer quality is based on the ratio between answers scores

and count of views (Kavaler and Filkov, 2018; Ravi *et al.*, 2014).

The difficulty of a question can not be explicitly observed, however Huna *et al.* (2016) approximated the difficulty with arrival time of first answer to a question. Other approaches were also used to estimate the difficulty of a question (Liu *et al.*, 2013; Sun *et al.*, 2018). Liu *et al.* (2013) gives an empirical distribution of questions difficulty on the mathematics and computer science questions on Stack Overflow. This distribution is an abnormal curve with two to five times more easy questions than difficult ones.

We consider two typical online user behaviors as Wu *et al.* (2016); Yang *et al.* (2014) characterized user behaviors into two main categories. In Yang *et al.* (2014), one category is who answer popular and difficult questions and the other one is who answer new and easy questions. Similarly Wu *et al.* (2016) classified all users into two answering strategies of answering easy or difficult questions. Their empirical study on different online communities showed the optimal mixing ratio is 63% of users answering simple questions for substantial growth of the community. Furtado *et al.* (2014) performed statistical clustering to find nine behavioral profiles using Stack Exchange data. More importantly this study showed that skill level is negatively associated with activity level and main contributors are more likely to be in the low activity class than in the hyperactivist class. This is consistent with the result from Yang *et al.* (2014).

3.2 Methods and Materials

3.2.1 Virtual Platform of Questions

Each question is represented as a basket with a width representing the question quality and depth as the sequential computing difficulty. The total difficulty of a

question is the product of width and depth. The width equals the number of bins in the basket. For configuration of the question basket, see figure 3.1.

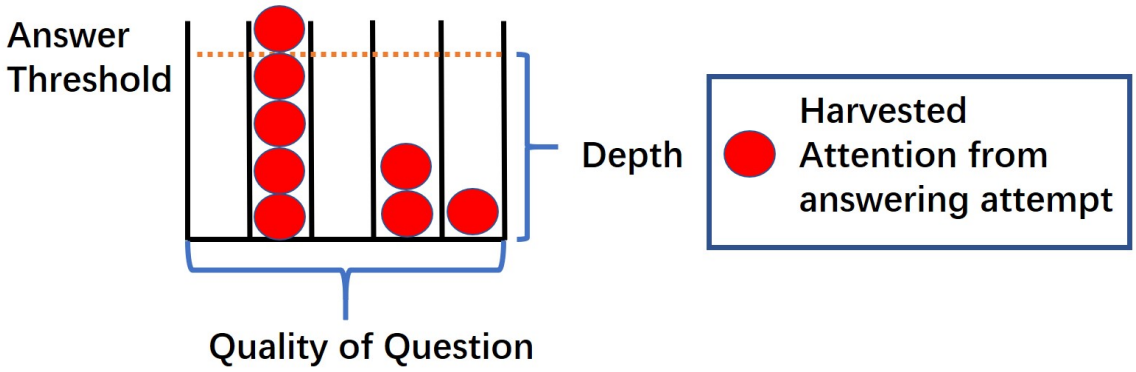

Figure 3.1: Concept of Question Answering and Question Difficulty

The number of bins reflects the quality of a question itself, which depends on how well the question is defined, its appropriateness related to the general topics of the community, and its tags. Li *et al.* (2012) investigated the entertainment & music category of Yahoo! Answers and gave distribution of question quality in a range of lowest 1 to highest 4. We generate the random variable B_q, the number of bins, based on the empirical distribution from their study.

As such when there are more bins, the question is not very specific or clearly stated. We define the question quality being lower when there are more bins like

$$Quality_q = \frac{1}{B_q} \tag{3.1}$$

where B_q is the number of bin for question q.

Differently in our model, 1 represents the highest quality score with bin size one and 0.25 is the lowest quality score with bin size four. A good quality question will allow users to easily understand the problem, convert the meaning and concentrate on one direction to derive an answer. So less bins means better quality of the question.

The difficulty to solve a question can be defined as the product of the depth and the number of bins.

$$D_q = \frac{Depth_q}{Quality_q} = Depth_q \times B_q \tag{3.2}$$

D_q is the difficulty to solve a question q. $Quality_q$ is the quality of question q and $Depth_q$ is the depth of question q. In the example of figure 3.1, the quality of question is 0.2 and the depth of the question is four thus the difficulty score would be 20.

Bad quality question could be misleading and increase the difficulty to solve while the true hardness of question itself could lay on the depth. Empirically the difficulty can be approximated with the arrival time of first answer to a question(Huna *et al.*, 2016). Other approaches were also used to estimate the difficulty of questions from Q&A community (Liu *et al.*, 2013; Sun *et al.*, 2018). We seem the study of difficulty distribution can imply the distribution of depth in our model.

3.2.2 Derivation of Answers

Attention is harvested and used to produce answers. Attention is represented by the red balls in figure 3.1. One unit of harvested attention equals to one red ball, meaning the amount of attention needed to solve one simplest and best quality question. Red balls are randomly placed in bins with equal likelihoods to solve a question. Only when the number of consecutive red balls in one bin exceeds the answer threshold, which is the depth of the question $Depth_q$, then there is one answer being successfully derived for the question. Similarly, multiple answers can be derived if the total number of consecutive balls in bins exceed multiple times of threshold $Depth_q$. In figure 3.1, only one answer was given to the question and the rest of two bins fail to derive an answer.

The principle of how to throw balls into different question baskets varies bu user types. We consider two typical users' behaviors: only answering difficult questions and only answering easy questions. The answer arrival time and the popularity of questions are not considered in our model. We assume the difference on question selection completely depends on the level of question difficulty. In our study, one type of users Y_1 are called difficulty lovers who prefer in answering difficult questions and the others Y_2 are called difficulty haters who prefer in answering easy questions.

Choice model is widely adapted to model online browsing and purchasing decisions (Nair, 2010; Park, 2017). Here we use choice model to calculate the probability of choosing a question q with difficulty score of D_q to answer given a user type. Let indicator variable $K_q = 1$ indicate that question q was chosen and latent variable K_q^* indicate the potential utility from choosing question q. K_q^* is purely determined by question difficulty D_q.

For Y_1 type users, K_q^* is positively related to question difficulty D_q

$$K_{q|Y_1}^* = \alpha + \beta D_q \tag{3.3}$$

For Y_2 type users, K_q^* is negatively related to question difficulty D_q.

$$K_{q|Y_2}^* = \alpha - \beta D_q \tag{3.4}$$

where parameter α represents the random effect on decision with individual- and time-specific variance and parameter β is the coefficient for difficulty. $\alpha = 0$ would suggest that no effect on individual difference and time variants when they are making choice.

The probabilities of choosing question q $P(K_q = 1|Y_1)$ when given user type Y_1 and $P(K_q = 1|Y_2)$ when given user type Y_2 are:

$$P(K_q = 1|Y_1) = \frac{e^{K^*_{q|Y_1}}}{\sum_i^Q e^{K^*_{i|Y_1}}} \tag{3.5}$$

$$P(K_q = 1|Y_2) = \frac{e^{K^*_{q|Y_2}}}{\sum_i^Q e^{K^*_{i|Y_2}}} \tag{3.6}$$

where Q is the total number of question.

3.2.3 Quality of Question with Answers

When the question is provided with answers, the total quality of question and answers together is U_q.

$$U_q = Quality_q \times Depth_q \times A_q \tag{3.7}$$

where A_q is the number of answers provided for question q. If no answer was derived, U_q would be zero. This quality of answered question equation 3.7 captures many features indicated in previous studies that high quality questions will more likely to attract high quality answers and more attempts to answer so that it increases the number of answers A_q(Baltadzhieva and Chrupała, 2015). Thus $Quality_q$ and A_q have positive impacts on the quality of question and answers pair U_q. Difficult questions are getting more quality answers as well based on the behavioral study of answering (Wu *et al.*, 2016; Yang *et al.*, 2014). It is important to distinguish the different definitions of difficulty in different studies. Wu *et al.* (2016); Yang *et al.* (2014) referred difficulty as true sequential computing difficulty to solve a question. While in reality difficulty to solve a question could due to bad quality question itself(no example, bad problem statement, wrong tag, etc) and sequential computing difficulty. Here in our model we consider the difficulty to answer a question comes from sequential computation difficulty modeled as depth $Depth_q$ and question quality $Quality_q$. In term of quality for the answers provided, one popular method of evaluating the answer quality is

based on the ratio between answers scores and count of views (Kavaler and Filkov, 2018; Ravi *et al.*, 2014).

Answers quality shall positively depend on the amount of attention and human computing power used to solve the questions. Therefore U_q directly depended on the number of balls in the basket N_q. Note that $Depth_q \times A_q \leq N_q$. It is reasonable to use $Depth_q \times A_q$ instead of N_q because if the same amount of attention spent on difficult and easy questions, there will be more waste of attention from failed attempts. The final quality of answers to an easy question would be higher if the amount of attention used to answer is the same between easy and difficult questions.

The overall average quality of answered questions across the community is $\bar{U}$:

$$\bar{U} = \frac{\sum_q^Q U_q}{Q} \tag{3.8}$$

Q is the total number of questions being answered across the community.

3.2.4 Return On Assets(ROA) Ratio for Answered Question

We define ROA as a way to measure the value gained from answering questions by paying one unit of attention. The related features to the question's value like the total number of answers provided, the total score of the answers, the time for highest-scoring answer to arrive, questioner's reputation and number of questioner's questions are monitored and shown to be important in Anderson *et al.* (2012). Therefore in our study, the value of answered questions with the corresponding answers V_q is determined by the total number of answers A_q, total number of balls in the basket of question N_q, and depth of question $Depth_q$.

$$\begin{cases} V_q = \beta_1 N_q + \beta_2 A_q + \beta_3 Depth_q & \text{for } A_q \geq 1 \\ V_q = 0 & \text{for } A_q = 0 \end{cases} \tag{3.9}$$

where $\beta_1, \beta_2, \beta_3$ are linear regression coefficients. The estimate of V_q is only defined when the questions have at least one successful answer. We define values for $\beta_1, \beta_2, \beta_3$ based on estimations from other studies. From Anderson *et al.* (2012) the value of β_1 equals to 0.83, which composes of regression coefficient for sum of answer scores 0.47, regression coefficient for number of comments on highest-scoring answers 0.19 and regression coefficient for number of comments on highest-reputation answerer's answer 0.17, since the amount of attention spent on one question will reflect by the sum of answer scores and number of comment. The value of β_2 equals to 0.61, which means the regression coefficient for the number of answers. Depth of a question contributes to the difficulty of a question, the length of the best answers and the comments of answers for sequential development of an answer thus $\beta_3 = 0.96$, which is the sum of regression coefficient for length of highest-scoring answer 0.38, coefficient of time for highest-scoring answer to arrive 0.22, and coefficients for number of comments 0.36.

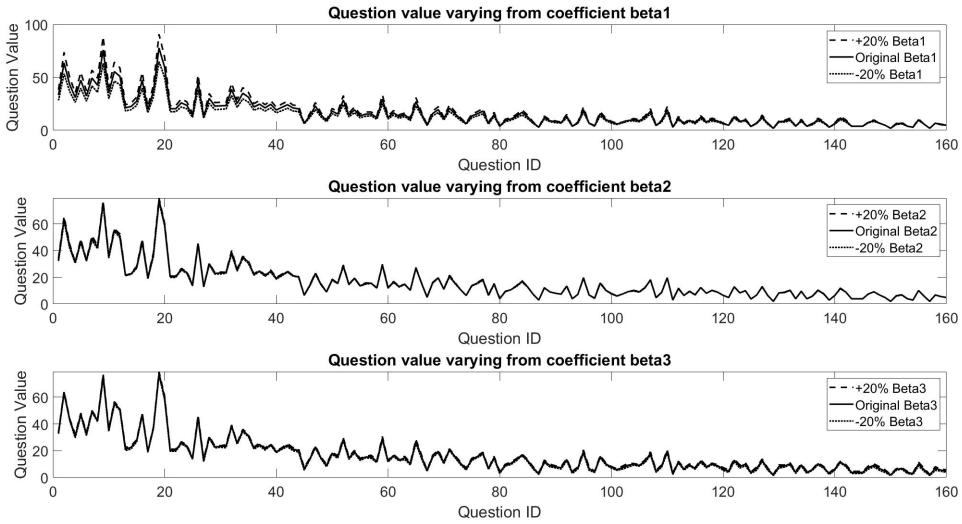

Figure 3.2: Sensitivity Analysis of Question Values by Independently Varying Coefficients β_1, β_2 and β_3 Values Up and Down 20%

Using distribution of number of answers and number of balls from simulation, question value is plotted in the figure 3.2 with question ID as x axis value and question value as y axis value. In fact, coefficients β_2 and β_3 are insensitive to the question value while coefficient β_1 has only minor effect on value for most popular questions as shown in figure 3.2. Changing curves from coefficients β_2 and β_3 is too identical to original curve to be clearly observed.

The Return On Asset(ROA) of attention harvesting on an answered question q is

$$ROA_q = \frac{V_q}{N_q} = \beta_1 + \frac{\beta_2 A_q + \beta_3 Depth_q}{N_q} \tag{3.10}$$

The overall ROA for the community $R\bar{O}A$ can be computed as

$$R\bar{O}A = \frac{\sum_q^Q V_q}{\sum_q^Q N_q} \tag{3.11}$$

Q is the total number of questions being answered across the community.

3.2.5 Simulation Algorithm

Simulation algorithm proceeds as flowchart 3.3 presents. Total attention, modeled by total number of balls in the system M is the sum of attention from both of type Y_1 and type Y_2 users, namely difficulty lovers and difficulty haters. Total number of balls M is a model variable to study. Other variables to study are r_d, a proportion of difficulty lovers type Y_1 among all population, and preferential coefficient of difficulty β in equations 3.3 and 3.4. At the beginning of simulation, system sets up three variables M, β, and r_d along with other constant parameters listed in table 3.1. Next, probabilities of balls assigned to each question by user types are discussed in previous section. After the assignments of balls to questions, we individually model the ball placement in the question basket that balls are randomly falling in different bins with even likelihoods. Finally, we summarize the metrics and compare contributions from two types of users.

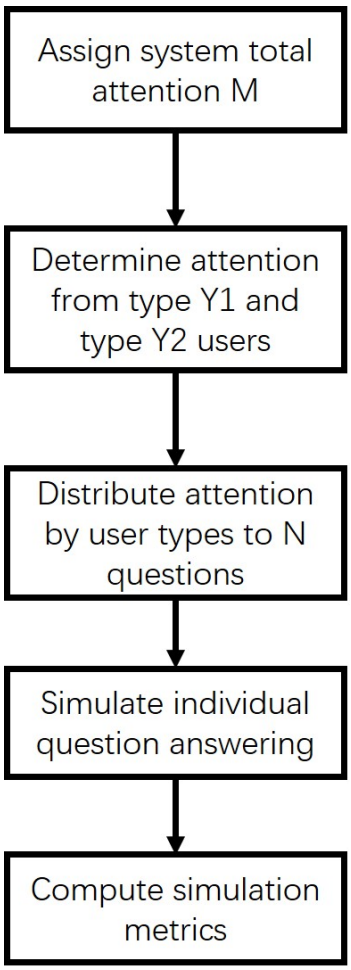

Figure 3.3: Flow Chart

3.3 Result

We simulate two identified answering strategies on the virtual community of questions and answers. We perform three special case studies 3.3.1,3.3.2 and 3.3.3 by varying one of three study parameters at a time. The parameter values used as baseline model, are listed in the table 3.1. The specific variations, made during special case studies, were explained in the corresponding subsection.

Table 3.1: Model Parameters

Parameter	Value	Case Study	Ref
N Total number of questions	1000	None	Assumed
M Total number of balls	5000	3.3.1	Assumed
r_d Proportion of difficulty lovers Y_1	0.37	3.3.2	Wu *et al.* (2016)
β_1 Regression coefficient for attention draw	0.83	None	Anderson *et al.* (2012)
β_2 Regression coefficient for number of answers	0.61	None	Anderson *et al.* (2012)
β_3 Regression coefficient for question depth	0.96	None	Anderson *et al.* (2012)
α random effect on decision	0	None	Assumed
β coefficient for difficulty	1	3.3.3	Assumed

The question quality was generated in four different levels based on the empirical result from Li *et al.* (2012) with 1 as the highest value of quality score for the question and 0.25 as the lowest. Four possible values for quality of question are 0.25, 0.33, 0.5, and 1.

Depth is generated by an approximation introduced in Liu *et al.* (2013) from 1-100 scale into 1-7 scale. Depth represents the true sequential computing difficulty of a question.

In the figure 3.4, the top left plot shows the histogram of question difficulty distribution. Most of the question is easy to medium level. At the top right of figure 3.4, we present the quality of answered questions and the corresponding contribution of balls from two separate user types, who are in favor of answering difficult questions and who are avoiding answering the difficult questions. As it shown, both types of answering behaviors contribute to high-quality questions and answers but difficulty lovers are contributing more to achieve the same level of high quality than difficulty haters. The plot at bottom left also shows the similar trend but in term of value created for question and answers. Although difficulty lovers are the contributors for high-value and high-quality questions, their ROA scores are lower than

difficulty haters, which means that they didn't utilize their attention as efficient and profitable as difficulty haters. Solving difficult questions requires more attention and effort, which drives difficulty lovers to spend more on difficult questions to achieve the same level of quality and value as for the easy questions. In the bottom right of figure 3.4, slightly lower ROA scores for questions with high number of balls means that, difficulty lovers spend too much attention on answering difficulty questions and these overly answered questions have lower ROA scores. If attention was put in other higher ROA questions, more value could be created.

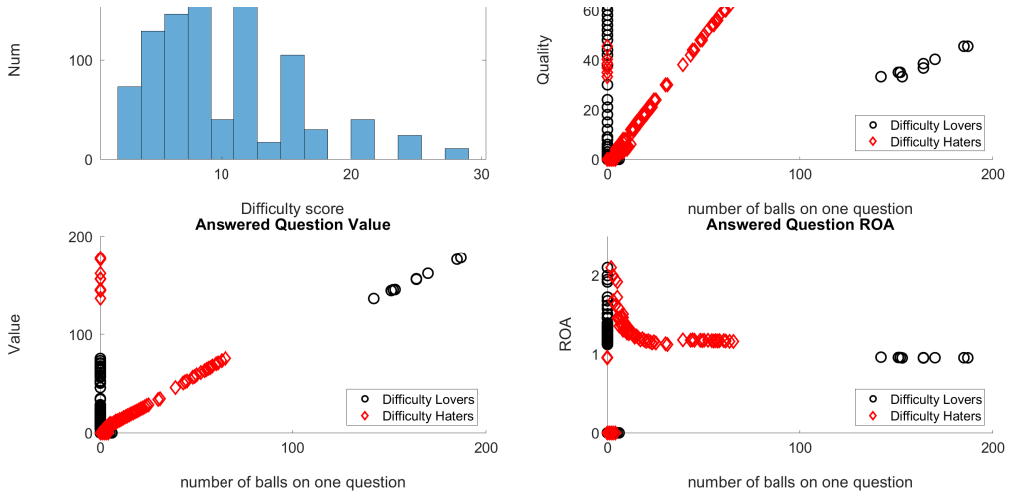

Figure 3.4: Distinct Patterns of Contribution by Two Users Types

3.3.1 Case Study 1: Variant Number of Balls

We increase M, the number of balls thrown per simulation, from 1000 to 40000. We see the percentage of questions being successfully answered increased as well as the overall average quality and value of an answered question. In the figure 3.5, the trends of three mentioned metrics match with our expectation, while the plot at the bottom right shows the ROA ratio is approaching to a steady level after a sharp

increase.

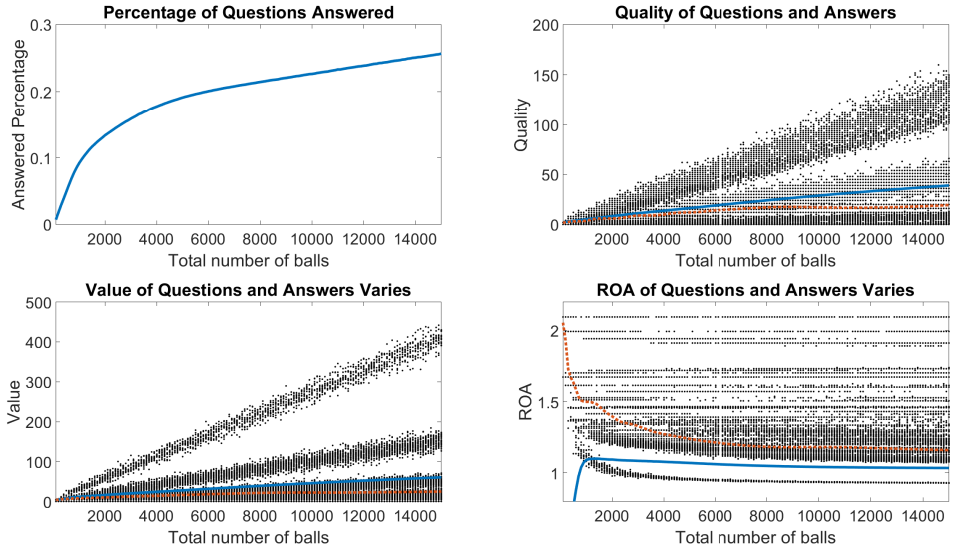

Figure 3.5: Dynamics of Increasing Number of Ball Participating(Blue Solid Line for the Average across the Community and Orange Dot Line for the Median across the Answered Questions)

The increase of answering percentage is nonlinear and is slowing down with a smaller slope. The orange dash line is the median value among answered questions and the blue solid line is the average value across the whole community. At the bottom right of figure 3.5, we see the median of ROA among all answered questions is decreasing. While we increase the total number of balls, the marginal value per answer from answered questions decreases to a steady level and less number of questions are unanswered. Median value of ROA is decreasing until an equilibrium was reached between newly answered questions with higher ROA scores and overly answered questions with ROA scores lower than 1. The overall average ROA shown in blue line lays in average ROA scores between two sets of questions, easy ones and difficult ones. Average ROA is lower than the median value of ROA in orange because of the balls

wasted into unanswered questions.

3.3.2 Case Study 2: Variant Ratio between Difficulty Lovers and Difficulty Haters

In the second case study, we look into the mixing ratio of difficulty lovers r_d in a population. By changing the proportion of balls for difficulty lovers from 0.1 to 1, we assign the balls thrown by difficulty lovers from 10% of the total number of 5000 balls to 100%. While we are increasing the proportion of balls for difficulty lovers, the balls left for difficulty haters are decreasing correspondingly.

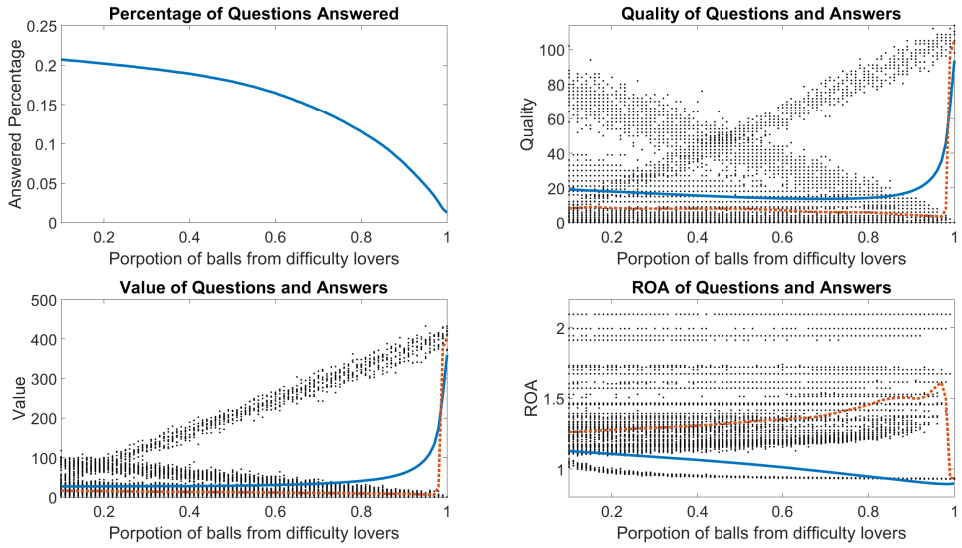

Figure 3.6: Dynamics of Changing Ratio between Difficulty Lovers and Difficulty Haters(Blue Solid Line for the Average across the Community and Orange Dot Line for the Median across the Answered Questions)

The percentage of questions being answered was decreasing in figure 3.6 while the total number of balls thrown by difficulty lovers increasing. Answering strategy is narrowing down to the minority of the questions with high difficulty and difficulty lovers don't contribute as much as the difficulty haters to the general question with

easy to medium level of difficulty. Also the value and quality of the questions are slightly decreasing toward the end and have a jump when $r_d = 1$. At the point of $r_d = 1$, the number of answered questions quickly slumps because of insufficient balls for general easy questions. Thus the questions answered are all by difficulty lovers and all the balls contribute to high difficulty questions. High concentration of ball distribution on a few highly difficult question when $r_d = 1$ causes a soaring value and quality among those highly difficult questions. However, the number of questions being answered dropped to lowest. In term of efficiency, we see from the bottom right in figure 3.6, that average ROA in orange color is decreasing when we are increasing the proportion of difficulty lovers. The median of ROA value across the questions being answered was increasing before approaching to the end, meaning that difficult answered questions have higher average ROA in general compared to easy answered questions. When approaching to the extreme state of $r_d = 1$, we see the average and the median of quality, value and ROA jump to the level of difficulty lovers cluster.

3.3.3 Case Study 3: Variant Preferential Coefficient of Difficulty β

Preferential coefficient of difficulty β in equations 3.3 and 3.4 is measuring how much the users value the difficulty while making decision to pick questions. Higher the coefficient β is, the more distinguished difference between the likelihoods of choosing highly difficult questions and very simple questions is. If β equals to zero, then it means that choosing a question is random and the difference in difficulty was not considered.

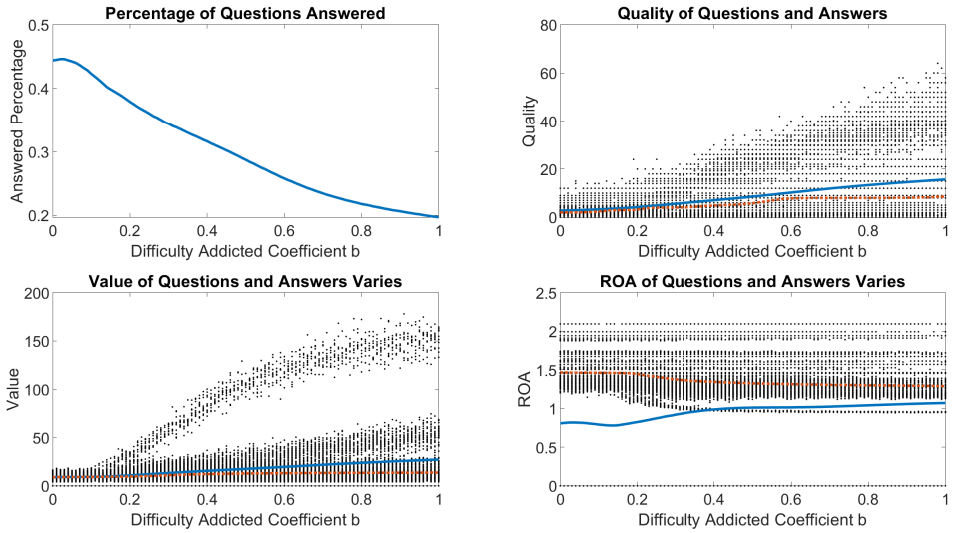

Figure 3.7: Dynamics of Changing Difficulty-based Coefficient b(Blue Solid Line for the Average across the Community and Orange Dot Line for the Median across the Answered Questions)

As the bottom left and top right of figure 3.7 shown, increasing the coefficient β will result in more attention paid in very difficult and very easy questions while this increase is scarifying the overall percentage of questions answered. The difficulty level of questions being answered is heading toward two extreme cases that either they are very tough or they are very simple. In average, quality and value are both slightly increasing. The average ROA is also slightly increasing meaning that, less balls have been wasted on the failure of producing answers.

3.4 Discussion

This project aims at measuring the efficiency of paying attention to answer questions, and analyzing contributions from different answering preferences in the difficulty level. We compute a Return On Assets(ROA) ratio to measure the profitability of attention used into creating value of question and answers by dividing the overall

value of answered question with total attention quantified as the number of balls in the question basket. Our algorithm simulates a virtual community under two identified and typical answering behaviors from Yang *et al.* (2014); Wu *et al.* (2016) which are interested in only difficult questions or only easy ones. More attention harvested results in higher average quality and value of a question and higher percentage of question answered. Three case studies find that the preference of answering difficult questions will have a lower efficiency of value creation. On the other hand, focusing on answering easy questions not only will significantly benefit overall percentage of questions being answered, but also show potential for improvement on overall efficiency. In this study, the limitation of our work is that the users dynamics and other preferences of user behaviors are not in the scope of our research. Only considering two answering behaviors are a simplified assumption for general user behaviors. This study is lack of the dynamics from other users types and collaborations between user types. Behaviors are simulated at a whole community level in our model. We are unable to study time factor and individual interaction. In the next project, we will further expand the study of user attributes and capture individual behavior.

Chapter 4

HOW BEHAVIOR OF USERS IMPACTS THE SUCCESS OF ONLINE Q&A COMMUNITIES

4.1 Introduction

In ecology diversity of species plays a critical role to maintain ecosystem's structure and function. It is known that ecosystems with less diversity are fragile(Elmqvist *et al.*, 2003). Species variety as a measure of system's robustness is a characteristic not only seen in ecology, but also in artificial systems like online communities as an analogy between diverse user behaviors on the online community and different species in an ecosystem. In Q&A online communities participants volunteer their time to answer questions. What makes Q&A online communities productive, meaning that a large number of their questions are answered in a satisfying way so that communities are beneficial in long term? This will depend on how participants channel their attention among the many different open questions. The conflict between information overload and scarcity of attention has been addressed by scholar for a long time (Simon, 1969; Eppler and Mengis, 2004; Lanham, 2006). The Q&A online communities is a practical example where effective challenging the attention of the community could lead to important community benefits.

In this paper we will use an agent-based model to explore the relationships between the diversity of behavioral strategies of agents on the performance of the Q&A community. This will help us to understand how incentive structures could impact the outcomes of Q&A communities. Studies on Q&A communities (Furtado *et al.*, 2014; Yang *et al.*, 2014; Samala, 2015; Wu *et al.*, 2016; Fang and Zhang, 2019; Wang

et al., 2018) identified and analyzed users' contributions, participation, answering strategy, and expertise to find distinct users groups. The difference between groups characterizes and emphasises on their different favors of difficulty and popularity on the questions to answer(Wu *et al.*, 2016; Yang *et al.*, 2014).

A large population scale study Hong and Page (2004) has shown that the more diverse a population is, the better capability of problem solving, compared to population of best-performing agents. However, there is little focus on what diversity of these user classes brings to the growth and accomplishment of online crowdsourcing community like Stack Overflow so that such online community becomes robust. Early work Alsina *et al.* (2015) measured the diversity as variation of the tenure of active users and compared to the score for answers across the community. Wu *et al.* (2016) showed a proper mixing ratio of strategy between answering difficult and easy questions helps the community to thrive. In this work, we analyze how decisions made at the individual's level can be captured and transformed into measures of the overall success of Q&A communities. Specifically we look into how the users' question selecting and answering strategies impact community's performance. This challenging task of aggregating and understanding emerging collective behavior and individual variability has also been pointed as a critical challenge in ecological problems (Levin, 1992).

In order to accomplish that, the first step is to set up a framework or environment with proper spatial and temporal scales and interaction mechanisms. Based on quantified factors from Liu *et al.* (2013); Li *et al.* (2012); Baltadzhieva and Chrupała (2015); Ravi *et al.* (2014); Huna *et al.* (2016); Sun *et al.* (2018), we develop a virtual environment of users participating in question answering activity on online community with distinct user behaviors and attributes observed in Wu *et al.* (2016); Anderson *et al.* (2012); Huberman *et al.* (1998); Kavaler and Filkov (2018); Yang *et al.* (2014);

Furtado *et al.* (2014).

With stochastic simulation and statistical interference, we explore the different contributions from each type of user group under their dynamic interactions. We further assess contributions by overall coverage of questions being answered, average quality and value of solved answers and accomplishment of average question difficulty on the community. Reaching and improving the above four criteria are the objectives we think online Q&A communities may want to accomplish. There are potentially many other ways to assess community's goals. Focusing on various objectives to community's success, we explore an multi-dimensional surface of feasible outcome under co-effect of question selecting and answering strategies. Our sensitivity analysis on these complex surfaces gives insight about how the owners and organizers of the Q&A community shall make trade-off on leverage to achieve objectives. This research also discovers relationship between question selection strategy and answering strategy and shows diversity of strategies crucial to achieve different objectives. Our study provides the groundwork to further studies on analyzing alternative designs for the development of the Q&A communities.

4.2 Model

The model describes the interaction of N_{user} agents with M questions. First we describe the construction of a virtual environment of questions including the definition of question quality, difficulty and answer formulation for a question in the subsections of previous chapter 3.2.1 and 3.2.3 for the quality of question and answers together. A virtual platform of questions consists of many baskets representing questions and new baskets will be constantly added into the environment. However, we assume there is no removal for the old questions. Thus the platform of questions is growing thought simulation time.

When a user spends attention on the question, the attention spent is modeled as balls placed into bins. All agents are assumed to randomly throw balls into the basket. An answer to the question is given when a bin reaches the answer threshold (see figure 3.1). Since a user throws balls randomly, more bins and higher thresholds require more effort to get answers to question.

Secondly we model users' activities on the virtual community described in 4.2.1. All users will have the same expertise level and the same amount of attention to contribute to the virtual environment. The model will explore how, given a limited amount of attention, different strategies to select and answer questions will impact the overall development of a Q&A community. At last, in the subsection 4.2.4 we illustrate steps to set up simulation and algorithm of simulation.

4.2.1 Diverse User Strategy

Each user selects questions to answer. Each user is assumed to have two important attributes which determine the strategies of which questions (s)he selects and how (s)he pays attention to answer those. Question selection strategy controls the preference in questions screening and picking potential questions to answer. After a user selects a number of questions to answer, how the user spends attention and focus to answer each question is governed by question answering strategy which determines the preference in answering which question first and last as well as how much attention and energy shall be used on trying to answer each question. In the following two sub-subsections we further illustrate the details of question selection strategy and question answering strategy. In the model we also assume independence between selecting preference and answering preference since, to our knowledge, there is no study showing there is correlation between two behaviors.

4.2.2 Question Selection Strategy

Users select a number of questions to consider to answer. The selection of this set of questions depends on the preference of the agent how to select. In Yang *et al.* (2014) one category of agents is owls (who prefer selecting popular and difficult questions to answer) and the other type of agents is sparrows (who are in favor of selecting new and easy questions). Classification of strategies about question selection preference for simply tasks versus challenging tasks is also adopted in Wu *et al.* (2016). During our study, we focus on two question characteristics: popularity and difficulty. We consider five independent strategies of question selection which are named easy, difficult, popular, new and random. Random strategy represents no preference. We assume users in the study independently belong to only one category of five, although in the reality users may have a mix of different strategies. The difficulty D_q (3.2) can be approximated by ratio between the number of answers and number of views in the reality(Huna *et al.*, 2016). If users have a preference to select difficult questions, the probability of a question q being picked $Prob^q_{difficult}$ by user is positively based on and ascended by absolute level on question difficulty D_q.

$$Prob^q_{difficult} = \frac{D_q}{\sum_i^Q D_i}\,. \tag{4.1}$$

Q is the total number of available questions. If users have a preference to select easy questions, the probability of questions being picked by user $Prob^q_{easy}$ is based on and descended by ranking on question easiness, the reciprocal of difficulty score.

$$Prob^q_{easy} = \frac{1/D_q}{\sum_i^Q 1/D_i}\,. \tag{4.2}$$

In our model, the number of views equally compares to the number of trials for balls N_q. Basket depth $Depth_q$ and bin number B_q, which difficulty D_q (3.2) composes

of, determines the likelihood of answering. A difficult question will have low ratio of answers per view. The nature of a question like sequential computing difficulty level and its own quality has been generated under distribution discovered by Li *et al.* (2012); Baltadzhieva and Chrupała (2015); Liu *et al.* (2013).

In realistic environment the number of views for a question and answers can approximate popularity from both answer seekers and question answerers. Here in our model we make assumption that the popularity of a question from selecting to answer is positively related and in same proportion to popularity of a question from answer seeking. Therefore, the number of balls in the basket represents the popularity of a question in our model under such assumption. It can be seen by everyone while selecting the questions to answer. The probability of questions being picked is based on and descended by absolute values on question popularity for users in preference of popular questions and reversely depended on popularity scores for users not in favor of popular questions but new questions. If users have a preference to select popular questions, the probability of a question q being picked $Prob^q_{popular}$ by user is positively based on and ascended by absolute level on number of balls in the question basket N_q.

$$Prob^q_{popular} = \frac{N_q}{\sum_i^Q N_i}\,. \tag{4.3}$$

The probability of a question q being picked $Prob^q_{new}$ by user in favor of selecting new questions:

$$Prob^q_{new} = \frac{1/exp(N_q)}{\sum_i^Q 1/exp(N_i)}\,. \tag{4.4}$$

We use exponential function $exp()$ to scale up the difference on weights between number of balls and do not need to worry the denominator becomes zero for a probability of selecting new questions.

For those in random category of question selection strategy, users will even likely and randomly choose from all questions. The probability of selecting any question q from random strategy user $Prob_{random}$:

$$Prob^{q}_{popular} = \frac{1}{Q}\,. \tag{4.5}$$

4.2.3 Question Answering Strategy

Given a pre-selected question set, a user will answer questions in a way determined by question answering strategy. Due to the issue of fragmentation of attention, it is normally the case that a user would not be able to pay long and continuous attention to solve questions especially for difficult questions requiring intense focus. We propose three answering strategies: "until one answer", "evenly", and "ROA".

"Until one answer" strategy is the most ideal case that user is able to prevent any distraction and concentrate on solving the question until (s)he derives an answer. The order of question answering is random.

We stochastically distribute attention or balls with equal probability among all selected questions which is the answering strategy of "evenly". The order of question answering is random.

At last, we propose a value estimation approach for questions and answers. This provides a measurement for gain on paying attention to answer called ROA (Return On Assets). Details of value estimation approach and calculation of ROA are given in the appendix 3.2.4. If a question has no answer, we can not compute ROA value and ROA will be defaulted to zero. The strategy of "ROA" will stochastically distribute attention based on ROA scores of all selected questions so that a higher ROA question will likely get more attention and also more likely be answered first. The likelihood of a question receiving attention is depended on the ROA score of a question. This

strategy tends to manage attention allocation and answering priority under the guide of ROA score so that users maximize the creation of value from answering.

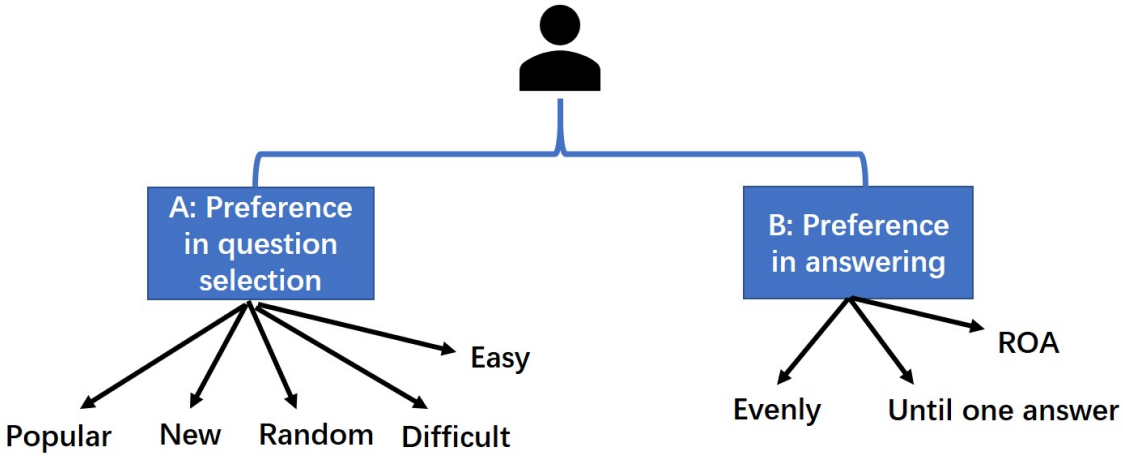

Figure 4.1: Structure of User Profile

In summary both "evenly" and "ROA" answering strategies gives maximum number of balls available to throw for each selected question. If assigned balls for each question run out, user will move to the next question until all balls are thrown or all selected questions are answered. If one answer is successfully derived, the remaining balls for the question will be added into the pool of assigned balls for the next question.

4.2.4 Model Flow Chart

In this section we describe the model of the whole system, namely how the generation of questions and answering of questions are connected. In Figure 4.2, a flow chart is presented that depicts the flow of actions during the simulation. For details about the parameter values and variable definitions we refer to the appendix B.1.

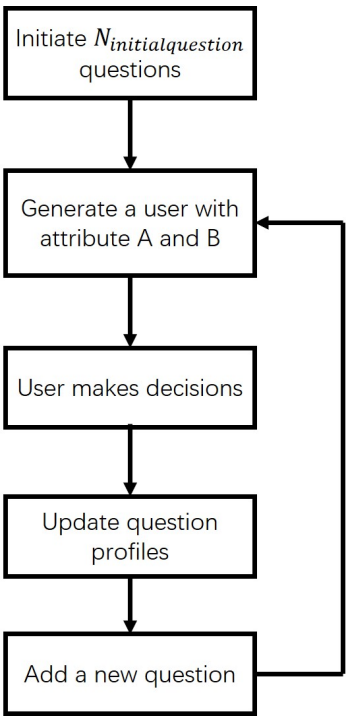

Figure 4.2: Flow Chart

At the start of a simulation we generate $N_{initialquestion}$ empty questions baskets with bins and depth. Initial number of questions $N_{initialquestion}$ is set to 10 for our model. Increasing or decreasing the value of this variable $N_{initialquestion}$ will not change long term dynamics but affects the speed of approaching saturation stage. A small value will create too much stochasticity at the early stage and a large value will slow down saturation speed. Number of bins and depth of the question basket are generated from empirical distribution from similar study on Q&A communities (Li *et al.*, 2012; Liu *et al.*, 2013). Quality of questions modeled as basket bin in our study is found to follow a gradually decreasing probability from low quality to high. Distribution of sequential computing difficulty modeled as basket depth is not normally distributed and has a very light tail toward difficult end. Then we start a sequence of steps that are repeated till the end of the simulation. We generate a user and assign attribute

A and B from user profile figure 4.1 to the user. This represents a user entering the space of questions during a certain time interval of interest. The user is defined by attributes A and B. With independent drawings those attributes are assigned. Percentage of each category is variable as we will explore in the model analysis the consequence of different distribution of attribute types. Based on attribute A the user will select the questions to answer. Based on attribute B a strategy is used to allocate the attention to answer the selected questions. Details of both are discussed in the previous subsections 4.2.1. We update the metrics how questions are answered, such as the number of questions successfully answered, calculate the value of the answered questions, the ROA and quality from user's contribution. We then add one new question into the community with a fixed rate of question introduction at each time step. Its number of bins and depth follow the same distribution as before. It is a new empty basket. If we are not at the end of the simulation we will continue with the next user.

4.2.5 Simulation Scenarios and Standards for Community Performance

First we discuss what determine the success of a Q&A community. Percentage of questions being answered on the community is good evaluation metric but it does not reflect the quality of answers and how likely it satisfies the asker's need. Let Q be the total number of questions in the observed simulation period and Q_s be the number of questions with at least one answer among all Q questions. The first standard P_s is evaluated by:

$$P_s = \frac{Q_s}{Q}\,. \tag{4.6}$$

Quality of answers(computation explained in appendix 3.2.3) shall be considered. In our model it is partially reflected by the number of answer provided to the question.

The second standard $\bar{QA}$ is given by:

$$\bar{QA} = \frac{\sum_i^{Q_s} QA_i}{Q_s} \,. \tag{4.7}$$

where QA_i is the quality of answers and question together for question i and question i must be answered already.

Naturally if a question is newly posted and attracts little attention and interest, having high quality answers to the question will not bring too much benefit as social good for the society. Ideally we want popular issues being addressed with high quality answers first than less interesting questions. A value estimation approach has been proposed in appendix 3.2.4 to quantify the long term value of one question and its answers. We consider this as another important statistical metric to look into while evaluating the success of a community. The third standards $\bar{V}$ is computed as:

$$\bar{V} = \frac{\sum_i^{L_w} V_i}{L_w} \,. \tag{4.8}$$

where L_w is the window size of question selection and defines the total number of questions that user can select to answer. It is set to 5 during our simulation. V_i is the value of answers created for question i. The definition of value V_i can refer to equation 3.9 in the appendix 3.2.4. If the question has no answer yet, the value is defaulted to zero.

Beside all of above standards, we believe on the community level, ability of solving difficult task is crucial as well. Research Wu *et al.* (2016) pointed out this ability increases the competency for online Q&A community, attracts more people to visit and increases more questions being asked. The average difficulty among all $\bar{D}$ is important metric to value competence of community as part of standard to success.

$$\bar{D} = \frac{\sum_i^{Q_s} D_i}{Q} \,. \tag{4.9}$$

D_i is the difficulty level for question i. Only answered questions will be considered into calculation of average difficulty thus it is different than the average difficulty of all questions on the community.

In conclusion, we look into four standards: percentage of questions being answered, average quality of answered questions, average value of selected questions and average difficulty of answered questions in our simulation as measurements of success.

Regardless of initial condition, whole simulation environment in term of average percentage of questions being answered will enter saturation stage after a warming up period when a balance between the newly arriving users and newly created questions forms. On average the percentage of questions being answered reaches a steady level when approaching to saturation stage shown in figure 4.3. We compare performance between scenarios and strategies after saturation. After examining the system and simulation result, we find sufficient warming up period is 50 time steps and all the results are collected from 51 to 150 time steps to guarantee the process reaches the steady stage. Since we introduce one user and one question per time step, Equally speaking, we stop simulation after we introduced the 150th user.

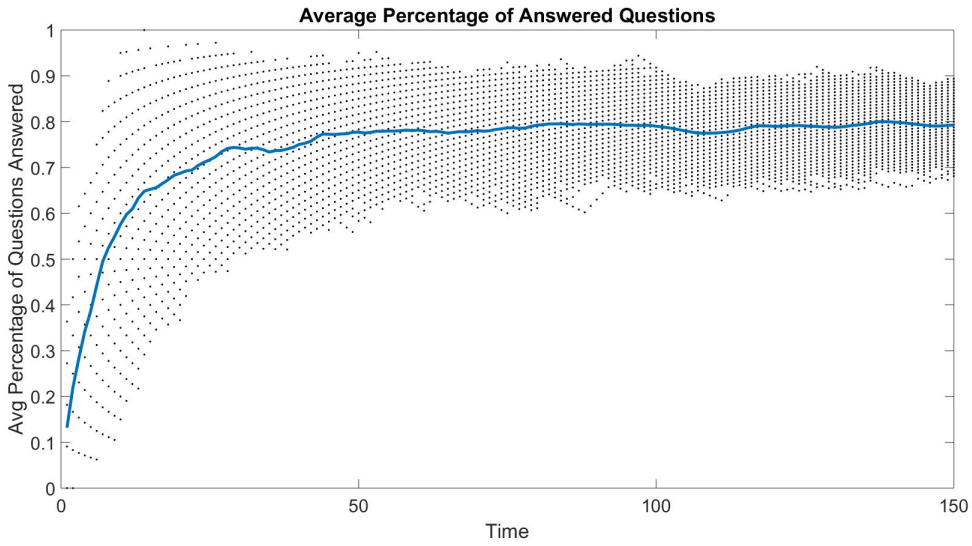

Figure 4.3: The Mean Percentage of Question Answered among 2000 Simulations: Black dots for One Simulation at One Time Step and Blue Line for the Simulation Mean at Each Time Step

The model parameters in the table B.1 are fixed across all simulations result shown in the main paper. We also study a few special and extreme cases that we change one or two parameter values from the table. The purpose of those special studies is to better understand and demonstrate model characteristics. Most of the results from special cases are in the appendix. The parameter values we are varying are clearly explained in the context.

We have five preference categories for question selection strategies: easy, difficult, popular, new and random. Category "random" can be alternatively replaced by a special case of equally mixing between categories "easy" and "difficult". We show such difference is trivial in the appendix B.2. This is one of the special simulation study that we perform to compare dynamic difference. We eventually only consider four question selecting strategies: easy, difficult, popular, and new for this part of question selection strategy simulation.

Disregarding the effect of random question selection strategy, categories easy and difficult can be named together as group of difficulty-biased no matter they are in favor of easy questions or difficult ones. New and popular categories can be named together as group of popularity-biased. For our interest of study, we present results for these two groups of popularity-biased and difficulty-biased in subsection 4.3.1. We conduct sensitivity analysis on mix of categories for different scenarios(defined in table 4.1) to explore the effect of group size on shaping the surface of variant proportion of mixture between difficulty-biased and popularity-biased groups. Dynamics of answering strategy is not within our scope, thus three answering strategies are equally and randomly assigned to users with probabilities in the table B.1.

We use a variable $ratio_x$ to represent the ratio between group size of popularity-biased and difficulty-biased. For instance if $ratio_x = 0.8$ that means 80% of whole simulated user population is falling into the group of popularity-biased and the rest 20% belongs to group difficulty-biased. The baseline scenario is when $ratio_x = 0.5$ and no group of random question selection, then we have equal population in groups of popularity-biased and difficulty-biased. Popularity dominated scenario is when $ratio_x = 0.8$ and difficulty dominated scenario is $ratio_x = 0.2$ without users of random question selection strategy. Proportions of three group for three scenarios are listed in the table 4.1. In reality, user group with identifiable behavior and preference is only a small portion and majority of users is behaving close to random strategy and no clear preference. Yang *et al.* (2014) finds two identified groups, the owls (who are experts and provide instructive answers to difficult and popular questions) and sparrows (who are highly active on the community, having a high reputation score and likely answering new and easy questions to accumulate the reputation scores), are no more than 23% of the total studied population on Stack Overflow. But they together make more than 87% answers. Therefore, the difficulty dominated scenario

is closer to the mixture in the reality while the majority of users from difficulty-biased group has no preference in choosing between easy and difficulty questions which act approximately like 50% in easy and 50% in difficult as we demonstrate before.

Table 4.1: Three Scenarios Setting

Scenarios	Proportions between three groups [Random, Popularity, Difficulty]	$ratio_x$
Baseline scenario	[0, 0.5, 0.5]	0.5
Popularity dominated scenario	[0,0.8,0.2]	0.8
Difficulty dominated scenario	[0,0.2,0.8]	0.2

4.2.6 Simulation Dynamics of Selection Strategy and Answering Strategy

At last, we look into dynamics of question selection strategy and answering strategy together. We have five question selection strategies and three question answering strategies. In total, we have 15 different combinations of selection-answering strategy pair. The first glance of the extreme cases, which only have one pair of selection-answering strategy on the whole virtual Q&A community, gives us ideas and insights about general behaviors of different pairs. Among 2,000 simulations, the average of number of answers, average increment of value and average increment of quality from each individual user are recorded and compared to 15 extreme cases.

More interesting, we would like to see how different strategies interact and collaborate on the virtual community. The second simulation framework has equal proportions of all question selection strategies and answering strategies. So speaking, the likelihoods of having selection strategies for random, easy, difficult, new and popular are all 20% and the likelihoods of having answering strategies for "Evenly", "Until One" and "ROA" are 33%, 33% and 34% respectively. Note that the assignment between selection strategy and answering strategy is independent.

We compare different answering strategies under different question selection strategies to find how one answering strategy will be more productive and suitable to another question selection strategy. Across 2,000 simulations we collect the mean of each estimation result under different combinations of answering strategies and selection strategies at each time step of simulation after time enters saturating stage. Results of 15 extreme cases with one single pair of selection and answering strategy and another simulation result with all possible pairs are in subsection 4.3.3.

4.3 Result

4.3.1 Dynamic of Mixture in Question Selecting Strategies

Let X be the proportion of users in the difficulty-biased group in preference of easy questions. Let Y be the proportion of users in the popularity-biased group in preference of popular questions. The conceptual chart of baseline scenario is presented in figure 4.4. For the other two scenarios, the group sizes of difficulty-biased and popularity-biased changes but the definition of variables X and Y remains the same. Variables X and Y are independent and are varied throughout simulations.

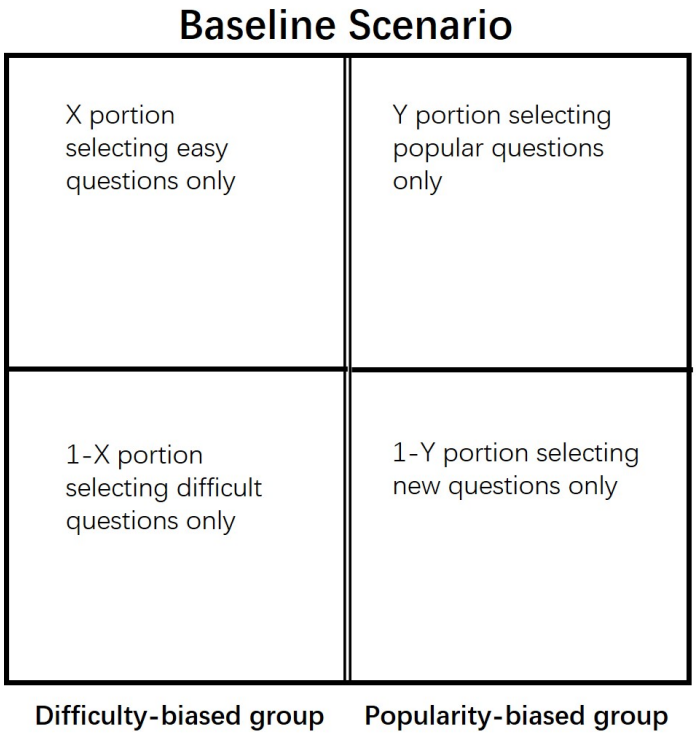

Figure 4.4: X Is the Proportion of Users in the Difficulty-biased Group in Preference of Easy questions. $1 - X$ Is the proportion of Users in the Difficulty-biased Group in Preference of Difficult Questions. Y Is the Proportion of Users in the Popularity-biased Group in Preference of Popular Questions. $1 - Y$ Is the Proportion of Users in the Popularity-biased Group in Preference of New Questions

In convenience of computation and result presentation, different values of variable X defines horizontal line of x axis which stands for the proportion of users in preference of selecting easy questions to answer from difficulty-biased group. Different values of variable Y defines vertical line as y axis stands for the proportion of users in the preference of selecting popular questions to answer from popularity-biased group. We separately vary and sample variables X and Y from 11 discrete points $[0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$. Discrete point 0 on y axis means no users has favor in new questions and 100% users in the group of popularity-biased are in

favor of popular questions. Similarly discrete point 0 on x axis means no users has favor in easy questions and 100% users in the group of difficulty-biased are in favor of difficult questions. Every discrete value of X has been coupled with different values of Y. We run 2,000 times simulation on each pair of different combination between X and Y. Mean of all simulation metrics among 2,000 simulations will be shown on 11 by 11 grid. We identify the maximum value per row and run t-test(normality tests satisfied) with significant level 0.05 to see how different compared to the row maximum the neighbor values along x axis would be. In detail the figures of results for different scenarios are given in the appendix B.3.1, B.3.2 and B.3.3.

In the following graphs 4.5, 4.6 and 4.7, all the lines connect to different shapes to show an optimal path determined by row maximums respectively for four different standards. Detail about optimal path can be referred in the appendix B.3.1. Lines indicate the optimal path that maximum values among horizontal axis moved along y axis, the proportion of preference in popularity. Circle indicates the global maximum point for average percentage of questions being answered. Star indicates the global maximum point for average value of questions being selected to answer. It is calculated as the total value gain from one user divided by the total number of questions selection. Maximizing the average value of questions being selected is the same as maximizing the total value of all selected questions gained from one round of user activity. Triangle indicates the global maximum point for average quality of questions being answered. Square indicates the global maximum point for average difficulty of questions among all questions in the community with difficulty level of unanswered ones defaulted to zero. It is computed as the average difficulty level among all solved and unsolved questions. Unanswered questions will have zero difficulty score. It captures the dynamics from question answering coverage and average level of difficulty among all answered questions.

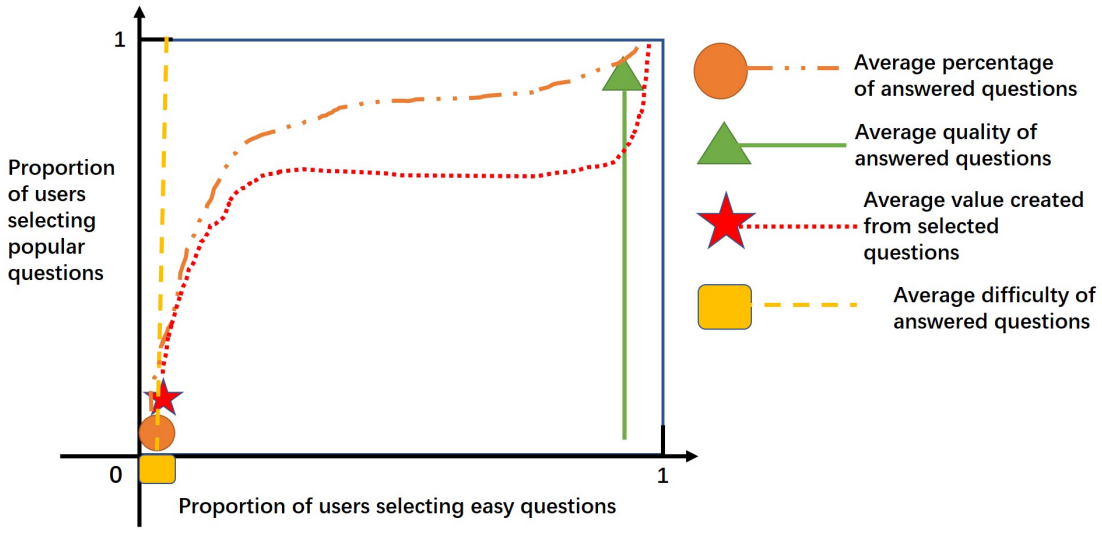

Figure 4.5: Row maximum paths and global maximum points between different standards for popularity dominated scenario (80% of popularity-biased and 20% of difficulty-biased)

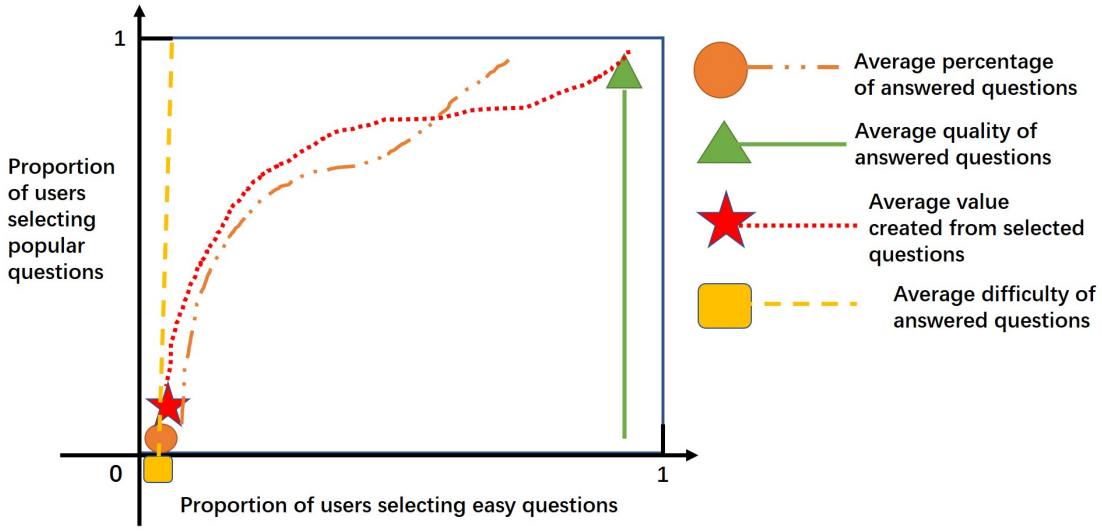

Figure 4.6: Row Maximum Paths and Global Maximum Points between Different Standards for Baseline Scenario (50% of Popularity-biased and 50% of Difficulty-biased)

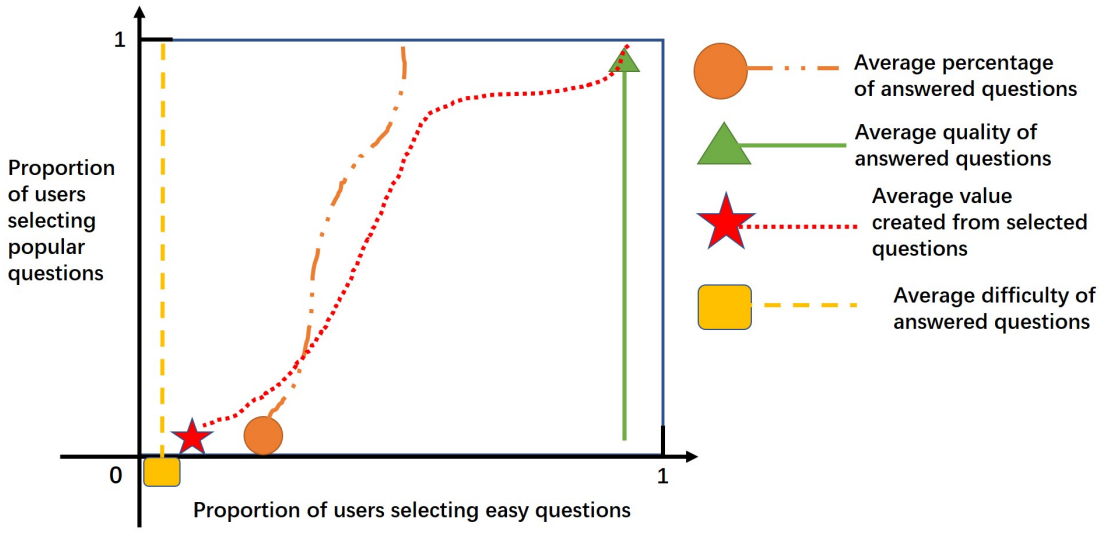

Figure 4.7: Row Maximum Paths and Global Maximum Points between Different Standards for Difficulty Dominated Scenario (20% of Popularity-biased and 80% of Difficulty-biased)

First thing to tell from all figures is that direction from right to left and from top to bottom maximizes the average percentage of questions being answered, average value created and average difficulty of answered questions for all three scenarios. Intuitively increasing the proportion of preference in new questions will maximize the average percentage of questions being answered. If there is enough coverage from users group in favor of new questions, easy questions will quickly and more likely be solved already than difficult ones. Thus more demand to answer difficult questions is needed. Orange circle is moving toward left to increase the proportion of preference in difficult questions when the proportion of preference in popular questions decreases.

Direction toward up right maximizes the average quality of answered questions. Global maximum point reached at 100% of population in preference in answering popular questions, because average quality of answered questions on the community directly depends on number of answers per question. More attention is spent on

solved and popular questions instead of unanswered and new ones will increase the number of answers per answered question. Therefore, more population in preference in answering popular questions, average total number of questions being answered decreases but number of answers per question increases. Easy questions require less attention to solve so if same amount of attention is put on difficult questions instead of easy questions, there will be less number of answers given to the questions. Therefore you can observe green line, which indicates the maximum path for average quality of answered questions always laying on right side at the point of 100% proportion of preference in easy questions.

Next, average value creation from selected questions moves in the similar way as average percentage of answered questions because only one answer can be given to one question by every user. In order to maximize the average value creation, strategy needs to maximize the total value created from answering selected questions. Value is only created when new answer is provided to question. It is the same goal as maximizing percentage of answered question. Thus the overall behavior of optimal path and movement toward global maximum point is similar to average percentage of answered questions.

In term of average difficulty of answered questions, 100% proportion of users in the group of difficulty-biased shall only focus on answering difficult questions to maximize the average level of difficulty. That is why the optimal path for average difficulty is laying at the left side border. One thing to point out is that calculation of average difficulty of answered questions is using total number of questions on the community to divide accumulated difficulty score from all answered questions. Ideal maximization strategy is to solve questions in priority of question difficulty. Global maximum point achieves when 100% proportion of users in the group of difficulty-biased chooses difficult questions and 100% proportion of users in the group of popularity-biased

chooses new questions.

The maximum paths of average quality of answered questions and average difficulty of answered questions were not affected by different setting of three scenarios. However, for average percentage of answered questions, more portion of difficulty-biased group squeezes the path along x axis, the proportion of preference in easiness, to tighter and sharpens the slope of the path along y axis, the proportion of preference in popularity. Similar stretching and squeezing effect are also observed in term of average value creation from selected questions. For the difficulty dominated scenario, both global optimal points of average value creation and average percentage of answered questions move away from left bottom corner and reach at 0.1 and 0.2 respectively. The interesting twisting and bending is due to the inequality of population size between two groups of difficulty-biased and popularity-biased. Presenting this dynamics of interaction and collaboration between this inequality group setting is one purpose of our simulation study.

All the analysis and results are based on heat maps on the discrete grid shown in figures B.3, B.6, B.4 and B.5 for baseline scenario. We perform analysis and interpretation for dynamic behavior around global maximum values. Figures B.7, B.9, B.8 and B.10 show corresponding analysis for the popularity dominated scenario and figures B.11, B.13, B.12 and B.14 show corresponding analysis for the difficulty dominated scenario.

4.3.2 Objective of Community Development

In this sub-subsection, we provide a potential implication of our finding about question selection strategy to evaluate different standards of community achievement. We compute the weighted score between all standardized scores of different standards. During the last sub-subsection we look at four standards to quantify the overall

performance of different question selection strategies on the community. If we give equal weight to each standard and plot the equally weighted score of four standards on the proportion map of difficulty and popularity, we produce the following contour plot figures B.15, 4.8 and 4.9.

Figure 4.8 shows contour map of equally weighted score of four standards: average percentage of questions being answered, average quality of answered questions, average value of questions selected and average difficulty of questions on the community. x axis is the proportion of population selecting easy questions to answer in the difficulty-biased group. y axis is the proportion of population selecting popular questions to answer in the popularity-biased group. The weighted objective score decreases along the y axis, the proportion of preference in popularity, toward more proportion of population selecting popular questions. We find the proportion of population selecting between easy questions and difficult questions does not affect the final objective score. Only population in preference of selecting new questions to answer is needed and ideal. The lighter area indicating the highest weighted score lays along the x axis of the proportion of preference in easiness and at the bottom of the low proportion of preference in popularity. The least attracting area in darker ink is when all population in the popularity-biased group is selecting only popular questions and all population in the difficulty-biased group is selecting only the difficult questions.

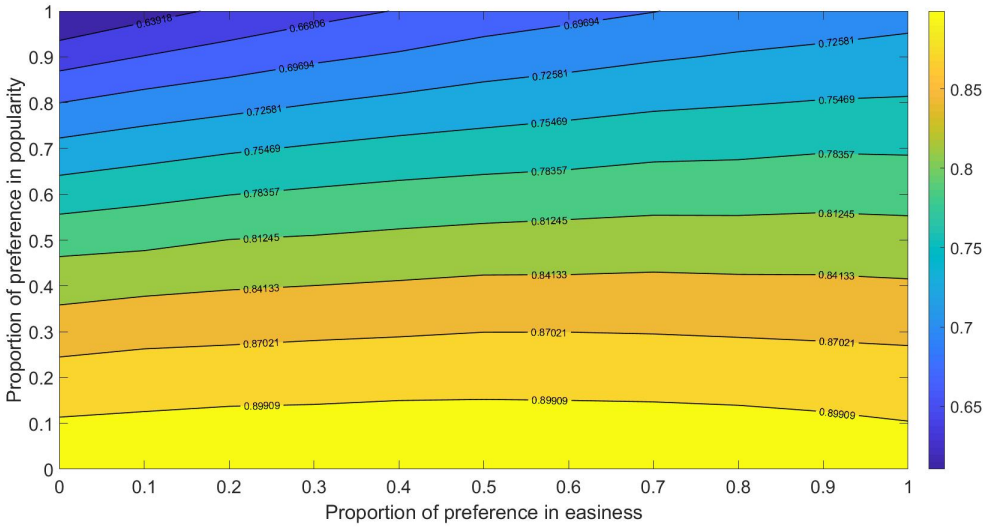

Figure 4.8: Contour Map of Equally Weighted Objective Score for Baseline Scenario

Among four standards, average value created from selected questions is the least sensitive one. Standardized score for average value of questions selected varies from 0.9 to maximum value 1. Thus the weight changing do not affect the overall contour map greatly for average value of selected questions. Average percentage of questions being answered and average difficulty of questions are the most sensitive two standards. That is why the maximum area from the contour map is dominated by the direction toward bottom left which is the maximizing direction of both average percentage of questions being answered and average difficulty of questions. Only for the average quality of answered questions, the maximizing direction of top right is in favor. The final weighted score of four standards on the contour map 4.8 shows the most attracting area is when no more than 10% population from group of popularity-biased is interested in selecting popular questions to answer while the mixture within group of difficulty-biased between choosing difficult questions or easy questions to answer does not impact the overall weighted score. Similar dynamic and observation can be found for the popularity dominated scenario and the contour map B.15 is

shown in the appendix.

However, for the difficulty dominated scenario we will see the contour lines no longer close to linear. In fact, we see nonlinear contour lines shown in graph 4.9 and extreme mixing proportion in the group of difficulty-biased(the proportion of preference in easiness close to 1 or 0) is no longer desirable.

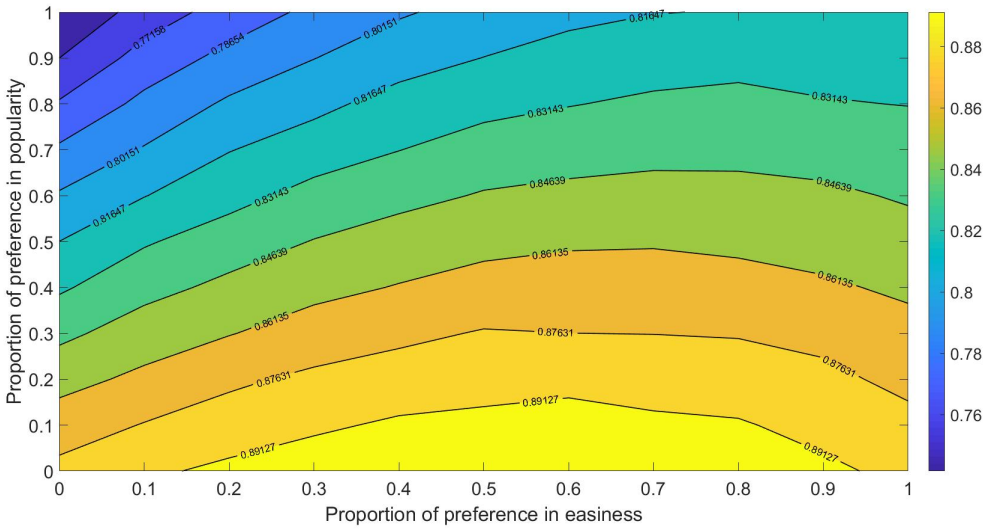

Figure 4.9: Contour Map of Equally Weighted Objective Score for Difficulty Dominated Scenario

Under the difficulty dominated scenario, majority of population is from the difficulty-biased group. Only 20% of whole population is from popularity-biased group. The proportion of mixing for difficulty-biased group becomes sensitive to the proportion of mixing for popularity-biased group. The maximum area in light color and contour lines in figure 4.9 is bending up depended on the distribution of question difficulty and interactions between selection strategies. In this case of scenario, diversity along the question selection between difficult questions and easy ones becomes very crucial to improve overall community performance. It is ideal to maintain a balanced mixture of population between preference in easiness and in difficulty at least when no more

than 80% of participants from popularity-biased group is concentrating on answering popular questions only.

Given different weight to each standard will certainly change the contour lines and maximum area. After all, equally weighting contour map is one example of implementation of sensitivity surface we explore with simulation and modeling. Long term goal for online community shall determine weight assignment for each evaluation standards. With certain weight combination, diversity of preference in easiness and difficulty may or may not be encouraged and optimal area could in favor of only one selection strategy between easy and difficult questions. If weight of average quality of questions answered is dominated, only selection strategy in easy questions is the attractive approach. While weights of average percentage of questions answered and average difficulty of questions answered are dominated, only selection strategy in difficult questions is ideal to use. Along y axis, the only optimal strategy is always 100% selecting new questions to answer. Including other different metrics as evaluation standard will also change the dynamics of contour map including maximum area and contour lines.

4.3.3 Finding about Pairing Different Answering Strategies with Ideal Question Selection Strategy

As we discussed in the sub-subsection 4.2.6, we observe outcomes from 15 extreme cases with single binding between selection strategy and answering strategy. The mean value is averaging the stochasticity of environment and dependence on previous state. The time series of the means are depending on the structure of question profile and both of individual answering strategy and selection strategy taken at the time step. Structure of question profile remain the same across all simulations thus the comparison on mean and variance of time series shall not be differentiated because of

structure of question profile. The final observation values shown in all figures are the mean of time series across 100 time steps from time point 51 to time point 150. 25 and 75 percentiles are captured and shown in red error bars. The variances of observation values stand for the variances of mean from 2,000 simulations and variance of question profile.

First, we look into average number of answers produced by individual user in the figure 4.10. Answering strategy of "ROA" produced greatly less number of answers when pairing with selection strategy of new. That is because "ROA" answering strategy always prioritizing(first answer and assign more attention to answer) question with large depth which is more difficult to answer only when the number of balls N_q in all question baskets is relatively small. When the number of balls N_q in all question baskets is relatively large and depth of question no longer dominates the ROA value of a question, "ROA" answering strategy tends to prioritize easy questions which have higher ROA values. That is why "ROA" answering strategy produced the most number of answers under selection strategy of popular. Answering easy question will generally create less value than getting an answer for a difficult question. This can be seen particular from the figure 4.12 that "ROA" answering strategy create the least amount of value under popular question selection strategy. All answering strategies under both easy and difficult question selection strategies are correspondingly getting similar amount of value even though the number of answers under easy question selection strategy is more.

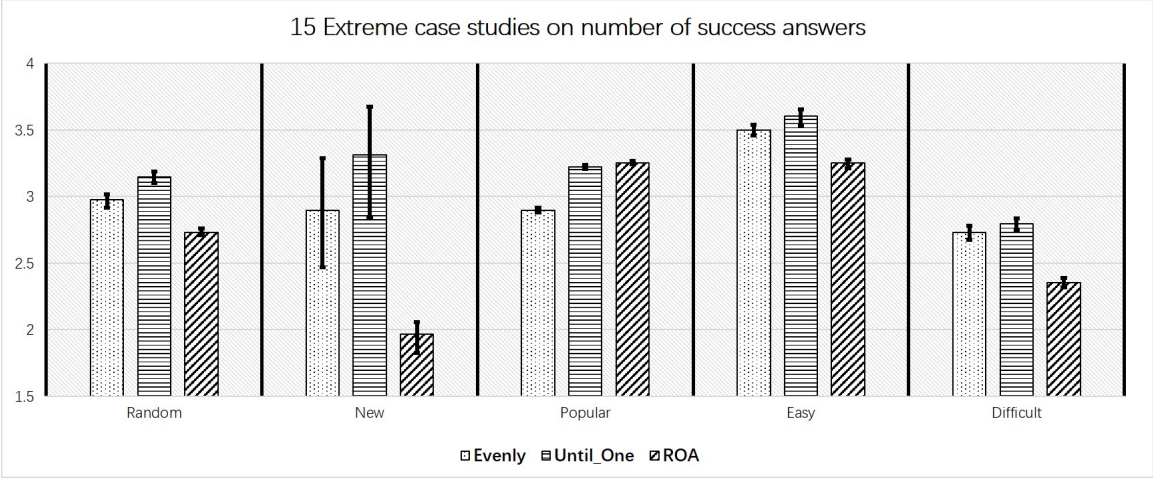

Figure 4.10: Average number of answers per individual user of 15 independent and extreme cases with single pair of selection and answering strategy and bar errors are 25th and 75th percentiles.

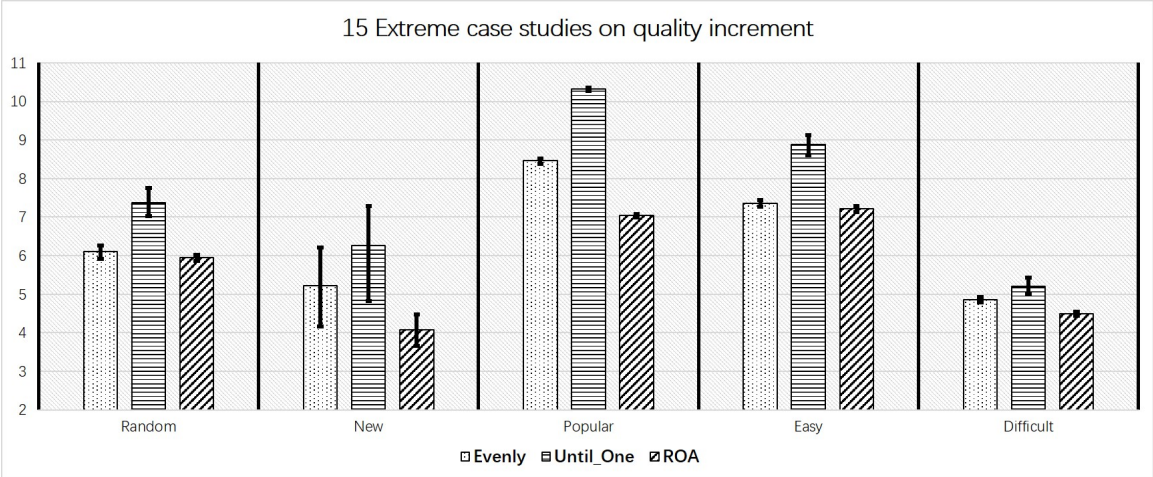

Figure 4.11: Average Increment of Quality per Individual User of 15 Independent and Extreme Cases with Single Pair of Selection and Answering Strategy and Bar Errors Are 25th and 75th Percentiles.

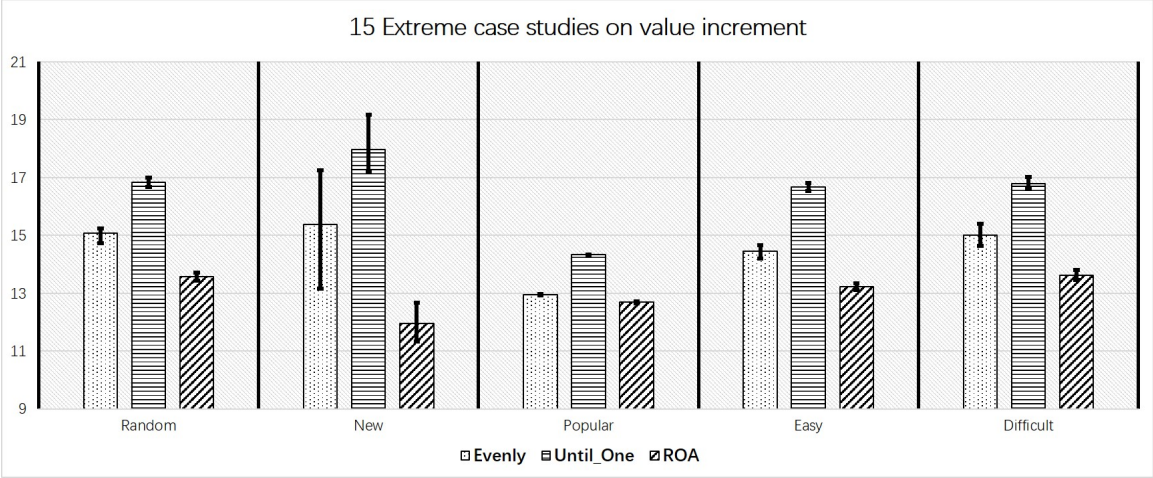

Figure 4.12: Average Increment of Value per Individual User of 15 Independent and Extreme Cases with Single Pair of Selection and Answering Strategy and Bar Errors Are 25th and 75th Percentiles.

Overall, "Until one" answering strategy is the best strategy to take and the second best is "Evenly" answering strategy. However, you will see during strategies interaction and collaboration in the following experiment, the least interesting answering strategy "ROA" becomes the second best answering strategy in term of average number of answers in figure 4.13 and average increment of quality in figure 4.15.

Next, we are not varying any model parameters but only observing the effect and performance of pairing between selection strategies and answering strategies. The setting of this focus study is using equal group size between five selection strategies: 20% of random, 20% of easy, 20% of difficult, 20% of popular and 20% of new. So speaking every user will have equal likelihood to adapt any one of five selection strategies. As for the answer strategy, it is equally distributed with 33% likelihood for "Evenly", 33% likelihood for "Until One" and the rest of 34% likelihood for "ROA".

The observation values are not normally distributed because of impact from question profile structure. We perform one-sample Kolmogorov-Smirnov test on all se-

quences of observation values. Null hypothesises, that the data in comes from a standard normal distribution, are rejected for all cases with confidence level $\alpha = 0.01$. Instead we use Mann-Whitney U-test, which has less assumption on the sample distribution, to examine the difference between group means instead of traditional t-test. Test results from Mann Whitney U-test imply the group mean differences between all pairs are statistical significantly different with all p values less than 0.001 expect for four cases. The exceptional cases are when question selection strategies are new and popular, group mean difference between "Evenly" and "ROA" answering strategies are insignificant for number of answers and increment of quality. p values from Mann-Whitney U-test for number of answers between "Evenly" and "ROA" are 0.0089 and 0.2051 respectively under new and popular selection strategies. p values from Mann-Whitney U-test for increment of quality between "Evenly" and "ROA" are 0.2647 and 0.9659 respectively under new and popular selection strategies.

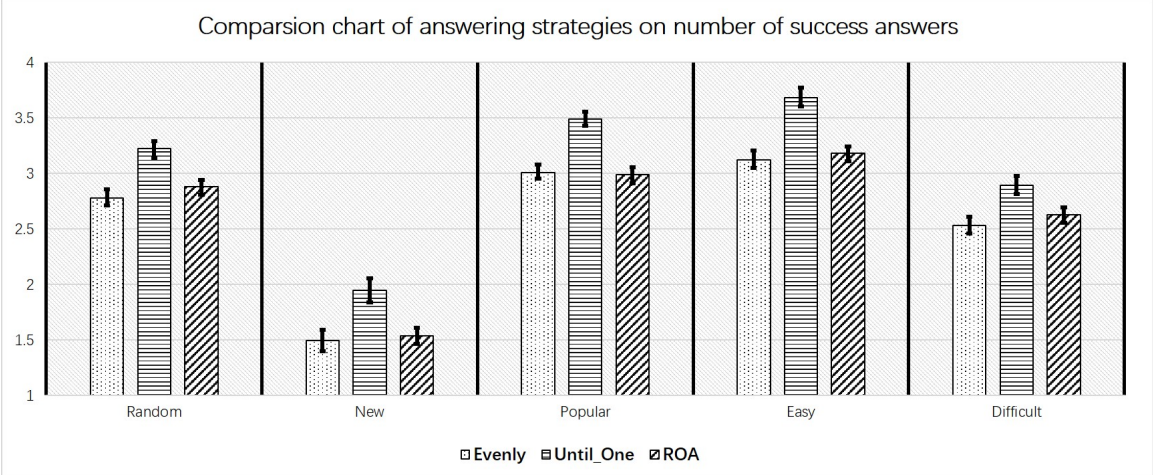

Figure 4.13: Average Number of Answers from Three Answering Strategies under Five Different Question Selecting Strategies with 25 and 75 Percentile Bars

In term of average number of answers being successfully produced to the selected questions, we observe that answering strategy "ROA" is producing more or equivalent

number of answers than "Evenly" answering strategy for all selection strategies while "Until One" is still the most productive answering strategy. Answering strategy "Until one" is outperforming because it better utilizes the residual of balls in the basket which are contributed by previous users but fail to form an answer. Ball residual is plotted in the bottom figures 4.14 by popularity level and 4.16 by difficulty level. Questions with top 25 percentile on difficulty level have been classified into difficult group with dot. Questions with bottom 25 percentile on difficulty level have been classified into easy group with triangle. The questions in between are called medium with circle. Shown in bottom figure 4.16, 25% of questions have depth less than or equal to 3 and contribute to majority of easy questions(difficulty level less than 15). The other 25% of questions have depth larger than or equal to 5 and widely reside in easy and difficult level. The rest of 50% of questions have depth of 4 and have difficulty level below 16. In summary, a difficult question(difficulty level larger than 16) must have a large depth. However an easy question could be a question with large depth but good question quality(less number of bins in the basket) or with small depth.

Shown in upper figure 4.14, all questions with high popularity have the same level of ROA scores and the difference between "ROA" and "Evenly" answering strategies becomes small because of dependence on ROA value of "ROA" strategy. It explains when selecting questions with high popularity, answering strategies "ROA" and "Random" is very close in producing number of answers because of similar ROA level among popular questions. When selected questions are new, ROA socre will most likely be defaulted to zero due to no answer provided. If all new questions have zero ROA scores, "ROA" answering strategy acts like "Evenly" to evenly distribute upper amount of balls to every question.

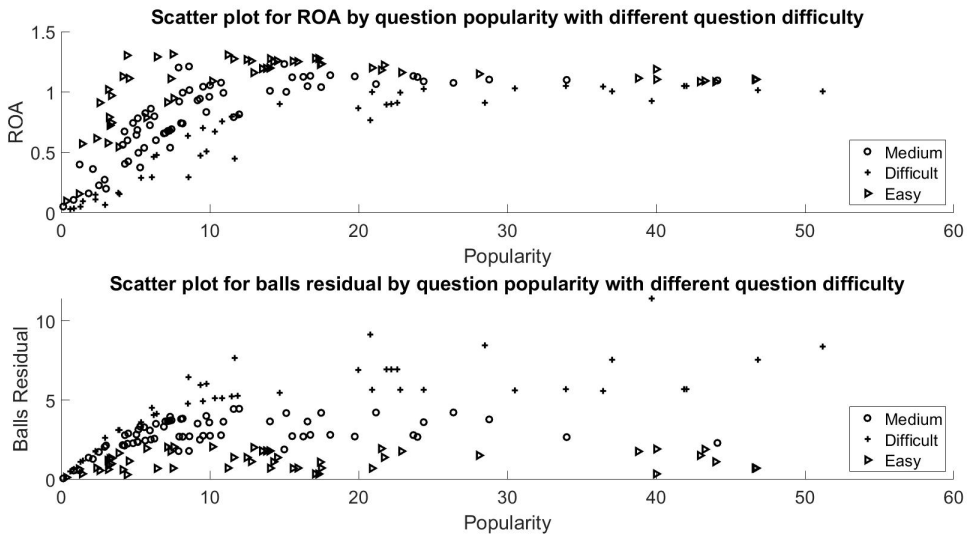

Figure 4.14: ROA and Balls Residual Plots along Popularity Level by Three Classes of Depth

Based on comparison on number of answers from figure 4.13, we can interpret the result of quality increment in figure 4.15 that more number of answers means more increment of quality. In average the question itself quality and depth is in the same level across all three answering strategies. In upper figure of 4.16, we can tell different difficulty levels and depth levels do not have differentiated level of ROA scores.

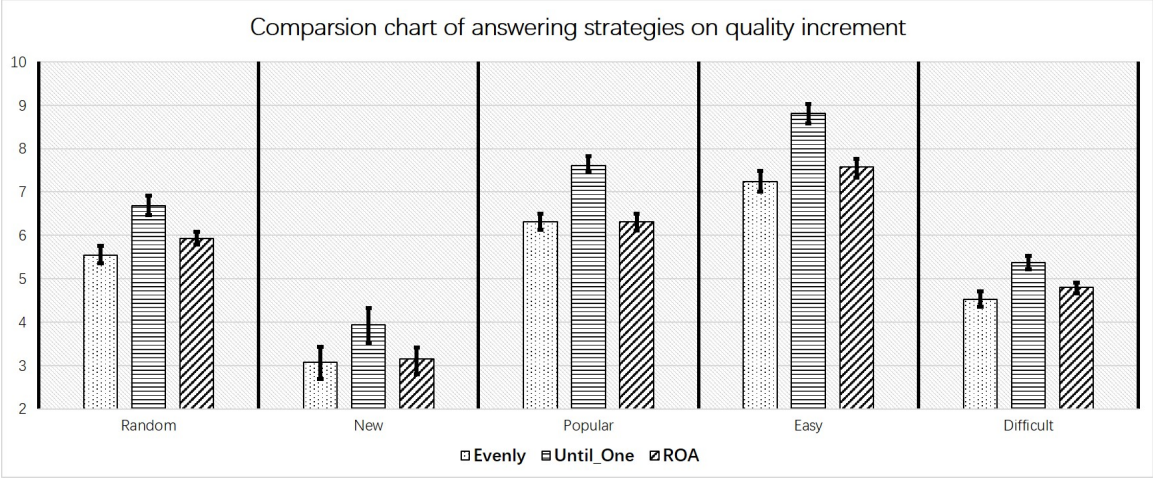

Figure 4.15: Average Increment of Quality from Three Answering Strategies under Five Different Question Selecting Strategies with 25 and 75 Percentile Bars

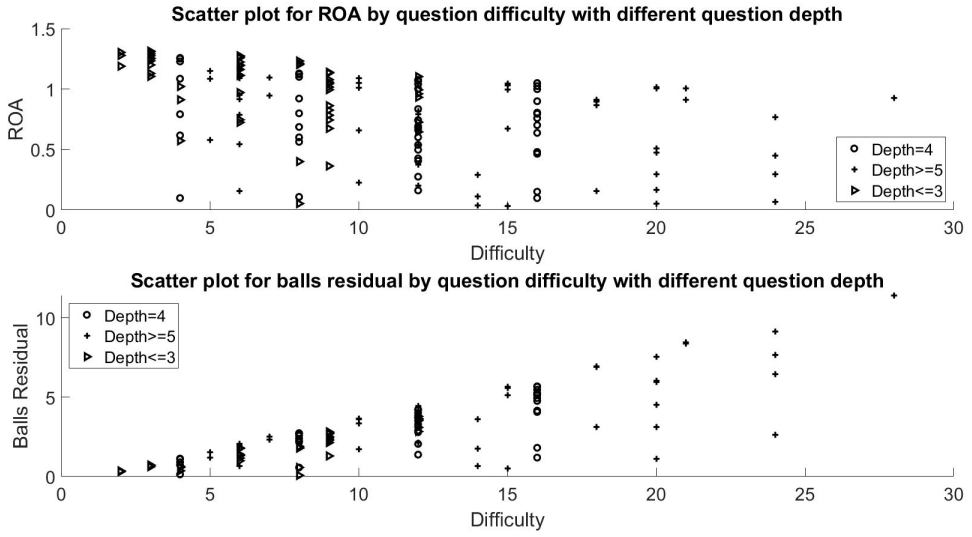

Figure 4.16: ROA and Balls Residual Plots along Difficulty Level by Three Classes of Depth

At last we look into the plot of value increment 4.17 between selection strategies and answering strategies. Value of questions depends on number of answers, attention to the question and answers or its popularity, which we modeled it as number

of balls in the basket during our study, and the depth. Unlike quality of question, which is the product between number of answers, depth and quality of question itself, value of question has linear and addable relation between each components. Detail of computation equation can refer to appendix 3.2.4. In configuration, first answer to the question claims a bonus on value related to the depth of question besides the value collected from the answer itself and attention to solve the question. Such dynamics is rewarding to the strategy which can most likely solve unanswered questions and that answering strategy is "Until one". While ROA for the new questions defaults to zero, answering strategy "ROA" ranks the new questions last to solve and increases the upper limits of attention used to solve questions with positive ROA scores. "ROA" answering strategy has the least chance to claim the bonus of being first answer to question while "Until one" answering strategy has the best chance to provide first answer to unsolved questions. Such tendency is reflected on the value increment plot 4.17 that "Until one" is the best answering strategy given all selection strategies. "ROA" answering strategy is the worst scenario expect for the case that question selection strategy is new. When all selected questions are new, "ROA" answering strategy behaves the same as "Random" answering strategy. However, given the situation of new questions mixing with questions having ROA scores, "ROA" answering strategy prioritizes questions with answers and ROA scores and allocates resource more efficiently to answer easy questions first(from upper figure 4.14). Given that most of selected questions are new without answer, efficiency and prioritization from "ROA" answering strategy provide slightly more number of answers to create more significant value than "Random" answering strategy which evenly allocates the resource to solve all selected questions.

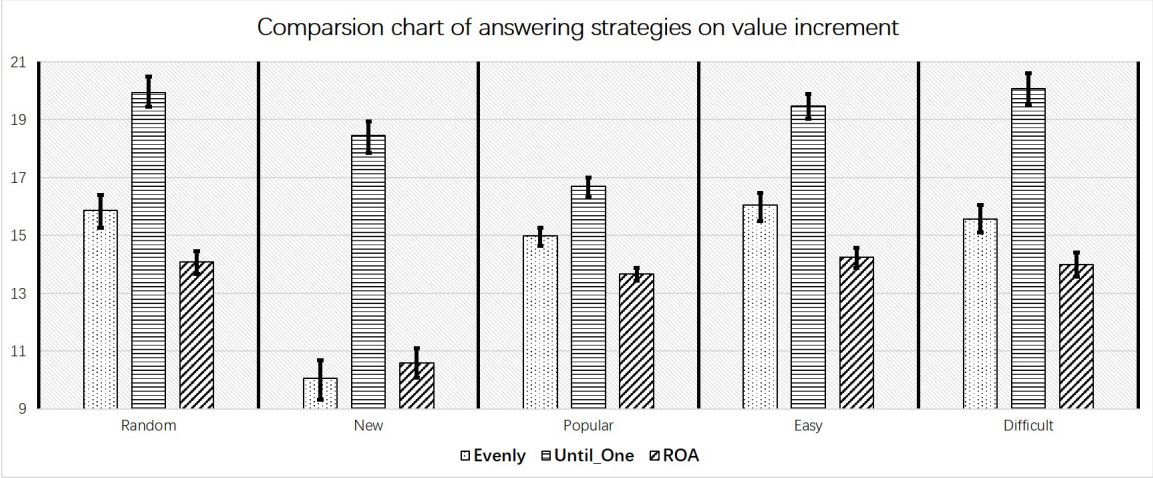

Figure 4.17: Average Increment of Value from Three Answering Strategies under Five Different Question Selecting Strategies with 25 and 75 Percentile Bars

In summary, first important point to make is that "ROA" answering strategy works better when collaborating with other answering strategies than with itself as we see big improvement on number of answers and quality increment from "ROA" answering strategy. Such improvement is caused by diversity of ROA scores and selected questions on the community shaped by diversity of question selection and answering strategies. While lack of diversity of selected question, dynamics of balls residual plays an important role to attract attention to answer with higher ROA score but relatively more difficult to answer due to the reason that newly answered question will have lower balls residual but higher ROA score.

Secondly, value increments and number of answers are all increased for "Until one" answering strategy except under new question selection strategy in the environment of diverse selection and answering strategies. It can be explained that diverse strategies create more balls residual. "Until one" answering strategy is effectively collecting the residual to form more answers. It is also true that implementation of single selection strategy and "Until one" answering strategy on the community can more effectively

solve questions with large depth so that quality increment from extreme cases with "Until one" is larger except for selection strategy of difficult. As mentioned before from bottom figure 4.16, difficult question must all have large depths(at least more than 4). If selected questions are all difficult, there are no distinguishing different on the depths between questions so "Until one" answering strategy will create more quality under the environment of diverse selection and answering strategies due to more number of answers.

Finally, under the environment of diverse combination of different selection and answering strategies "ROA" answering strategy is at least better than "Evenly" strategy on value increment and no worse than other two metrics when paring with new question selection strategy.

4.4 Conclusions

In the case of difficulty dominated scenario, we shows question selecting strategy of new is always preferred to popular after equally balancing all main community objectives and it would be optimal to have slightly more proportion of users in preference of selecting easy questions than selecting difficult ones. This finding of optimum area falls inline with empirical observation of group mixture from Stack Overflow under effect of its reputation rewarding mechanism. However, we discover that under other different mixtures of question selection groups community performance metrics could change dramatically. For instance, for the case of popularity dominated scenario, optimal mixture for average value of answered questions given proportions of users selecting popular and new questions has very steep slope of change and a jump with discontinuity near the point that 30% of users selecting popular questions and the rest of 70% users from the popularity-biased group is selecting new questions. The simulation result suggested the reputation rewarding system adapted by Stack Over-

flow is driving the community to a direction of improving overall successes. Some very active users under this rewarding system are pursing high reputation scores by looking for new and easy questions to answer and they are referred as sparrows(Yang *et al.*, 2014).

While paying continuous attention until question is solved or time runs out is the best answering strategy among all, in the reality it is hardly achieved because of fragmentation of attention. Under the existing of diverse question selection and answering strategies, the second best answering strategy is allocating attention to solve questions based on ROA scores of questions if the goal is to improve overall quality of answered questions on the community and total number of answers to the questions. Moreover it is also true for improvement of overall value, particularly when users is selecting new questions to answer as primary selection strategy. We discover the limitation of answering strategy "ROA" which is short of answering coverage for unsolved questions. The created value from "ROA" answering strategy is lowest in general even if it aims to maximize it because ROA scores sometimes are misleading and not reflect the dynamics of uncertainty.

The model makes arguable independence assumptions between question selecting strategies, answering strategies and answering order from question to question. On the real Q&A community, these independence may not be valid. In fact Yang *et al.* (2014) shows roughly 10%-22% of sparrows also belong to the set of owls. So speaking, User could adapt more than one selecting strategies and answering strategies. While there has no evidence to find dependent relationship or connection between question selection preference and answering behavior either as far as we know. Next, questions are linked to questions by relatedness and shared asker. Network of questions has been heavily studied. In the future work we would like to incorporate such study and discuss how the underlying network structure of questions changes the overall dy-

namics and result. Adding the question network structure will fundamentally change the affect of question selection strategy. It will also impact on the implication of answering strategies. Other potential work can be discussed is related to cost-effect analysis on changing mixture of question selection strategy.

REFERENCES

Albert, R., H. Jeong and A.-L. Barabási, "Error and attack tolerance of complex networks", nature **406**, 6794, 378 (2000).

Alsina, E. F., W. Rand and K. Lerman, "The success of question answering communities: How diversity influences ad hoc groups", Collective Intelligence pp. 1–4 (2015).

Anderson, A., D. Huttenlocher, J. Kleinberg and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow", in "Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 850–858 (ACM, 2012).

Bade, R. and M. Parkin, *Foundations of microeconomics* (Pearson/Addison Wesley, 2007).

Baltadzhieva, A. and G. Chrupała, "Question quality in community question answering forums: a survey", Acm Sigkdd Explorations Newsletter **17**, 1, 8–13 (2015).

Barabási, A.-L. and E. Bonabeau, "Scale-free networks", Scientific american **288**, 5, 60–69 (2003).

Brabham, D. C., "Crowdsourcing as a model for problem solving: An introduction and cases", Convergence **14**, 1, 75–90 (2008).

Elmqvist, T., C. Folke, M. Nyström, G. Peterson, J. Bengtsson, B. Walker and J. Norberg, "Response diversity, ecosystem change, and resilience", Frontiers in Ecology and the Environment **1**, 9, 488–494 (2003).

Eppler, M. J. and J. Mengis, "The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines", The information society **20**, 5, 325–344 (2004).

Fang, C. and J. Zhang, "Users' continued participation behavior in social q&a communities: A motivation perspective", Computers in Human Behavior **92**, 87–109 (2019).

Furtado, A., N. Oliveira and N. Andrade, "A case study of contributor behavior in q&a site and tags: the importance of prominent profiles in community productivity", Journal of the Brazilian Computer Society **20**, 1, 5 (2014).

Hong, L. and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers", Proceedings of the National Academy of Sciences **101**, 46, 16385–16389 (2004).

Huberman, B. A., P. L. Pirolli, J. E. Pitkow and R. M. Lukose, "Strong regularities in world wide web surfing", Science (New York, N.Y.) **280**, 5360 (1998).

Huberman, B. A., D. M. Romero and F. Wu, "Crowdsourcing, attention and productivity", Journal of Information Science **35**, 6, 758–765 (2009).

Huna, A., I. Srba and M. Bielikova, "Exploiting content quality and question difficulty in cqa reputation systems", in "International Conference and School on Network Science", pp. 68–81 (Springer, 2016).

Kavaler, D. and V. Filkov, "Determinants of quality, latency, and amount of stack overflow answers about recent android apis", PloS one **13**, 3, e0194139 (2018).

Kelley, T. M. and E. Johnston, "Discovering the appropriate role of serious games in the design of open governance platforms", Public Administration Quarterly pp. 504–554 (2012).

Khatib, F., S. Cooper, M. D. Tyka, K. Xu, I. Makedon, Z. Popović and D. Baker, "Algorithm discovery by protein folding game players", Proceedings of the National Academy of Sciences **108**, 47, 18949–18953 (2011).

Kim, J. B., P. Albuquerque and B. J. Bronnenberg, "Mapping online consumer search", Journal of Marketing research **48**, 1, 13–27 (2011).

Lanham, R. A., *The economics of attention: Style and substance in the age of information* (University of Chicago Press, 2006).

Levin, S. A., "The problem of pattern and scale in ecology: the robert h. macarthur award lecture", Ecology **73**, 6, 1943–1967 (1992).

Li, B., T. Jin, M. R. Lyu, I. King and B. Mak, "Analyzing and predicting question quality in community question answering services", in "Proceedings of the 21st International Conference on World Wide Web", pp. 775–782 (ACM, 2012).

Liu, C., R. W. White and S. Dumais, "Understanding web browsing behaviors through weibull analysis of dwell time", in "Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval", pp. 379–386 (ACM, 2010).

Liu, J., Q. Wang, C.-Y. Lin and H.-W. Hon, "Question difficulty estimation in community question answering services", in "Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing", pp. 85–90 (2013).

Nair, C., "Essays on online browsing and purchase", (2010).

Oulasvirta, A., "The fragmentation of attention in mobile interaction, and what to do with it. interactions, 12 (6), 16e18", (2005).

Park, C. H., "Online purchase paths and conversion dynamics across multiple websites", Journal of Retailing **93**, 3, 253–265 (2017).

Ravi, S., B. Pang, V. Rastogi and R. Kumar, "Great question! question quality in community q&a.", ICWSM **14**, 426–435 (2014).

Rutz, O. J. and R. E. Bucklin, "Does banner advertising affect browsing for brands? clickstream choice model says yes, for some", Quantitative Marketing and Economics **10**, 2, 231–257 (2012).

Samala, R. R., *Analyzing User Participation Across Different Answering Ranges in an Online Learning Community* (Arizona State University, 2015).

Shirky, C., *Cognitive surplus: How technology makes consumers into collaborators* (Penguin, 2010).

Simon, H. A., "Designing organizations for an information-rich world", Brookings Institute Lecture, September (1969).

Sun, J., S. Moosavi, R. Ramnath and S. Parthasarathy, "Qdee: Question difficulty and expertise estimation in community question answering sites", arXiv preprint arXiv:1804.00109 (2018).

Wang, C.-J., L. Wu, J. Zhang and M. A. Janssen, "The collective direction of attention diffusion", Scientific reports **6**, 34059 (2016).

Wang, J., Y. Luo, J.-X. Hao, L. Ding and R. Zhang, "Who are influential in q&a communities? a measure of v-constraint based on knowledge diffusion capability", Journal of Information Science p. 0165551518800411 (2018).

Wu, L., J. Baggio and M. Janssen, "The role of diverse strategies in sustainable knowledge production: e0149151", PLoS ONE **11**, 3, URL `http://search.proquest.com/docview/1776667078/` (2016).

Wu, L. and M. A. Janssen, "Attention dynamics in collaborative knowledge creation", arXiv preprint arXiv:1511.07616 (2015).

Yang, J., K. Tao, A. Bozzon and G.-J. Houben, "Sparrows and owls: Characterisation of expert behaviour in stackoverflow", in "International Conference on User Modeling, Adaptation, and Personalization", pp. 266–277 (Springer, 2014).

APPENDIX A

ADDITIONAL GRAPHICS FOR CHAPTER 3 PROJECT2

A.1 FIGURES

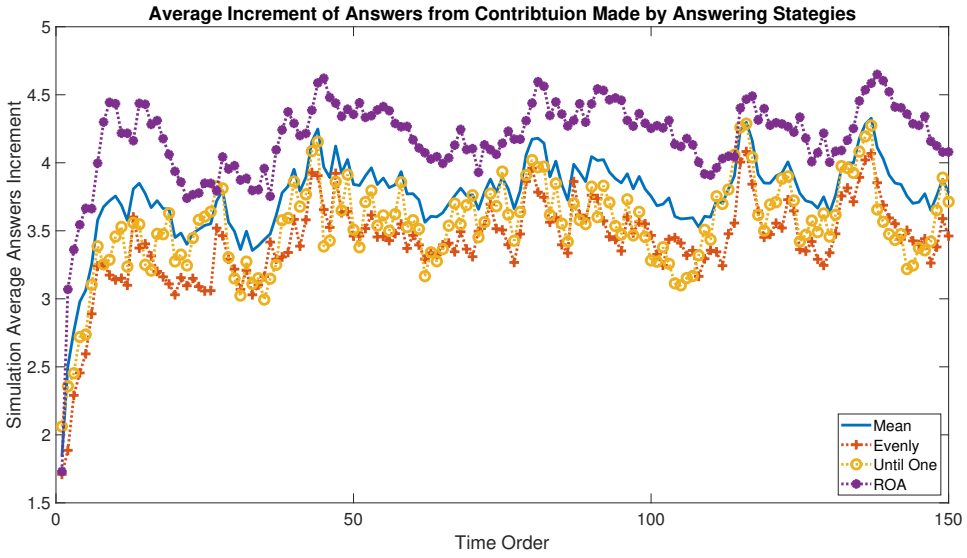

Figure A.1: Average Number of Answers Derived from Three Different Strategies at Each Time Step with Original Simulation Setting of 15 Balls on Hand

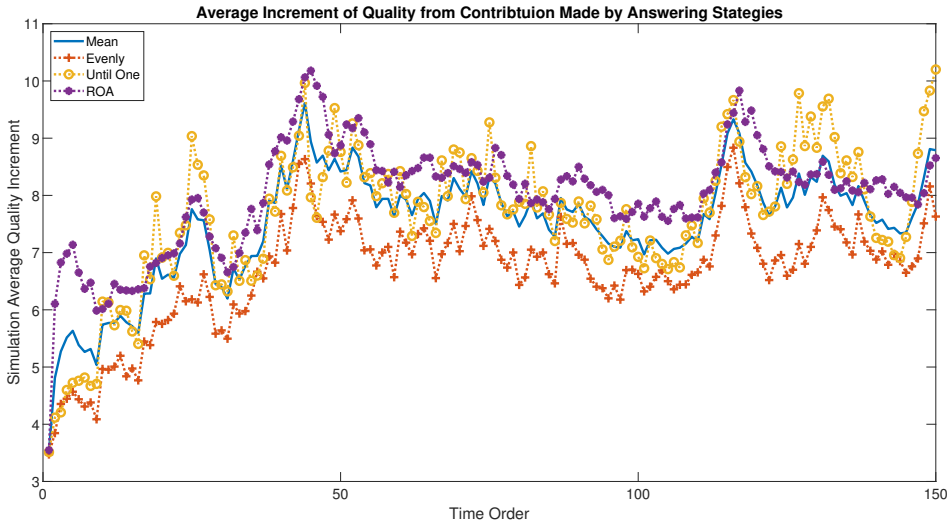

Figure A.2: Average Additional Quality Created from Three Different Strategies at Each Time Step with Original Simulation Setting of 15 Balls on Hand

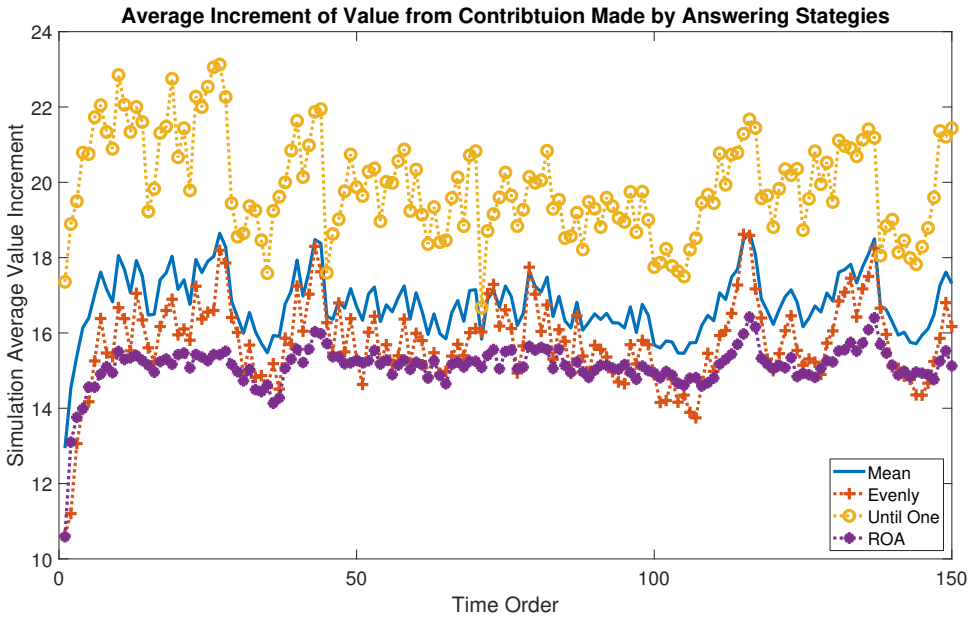

Figure A.3: Average Additional Value Created from Three Different Strategies at Each Time Step with Original Simulation Setting of 15 Balls on Hand

APPENDIX B

ADDITIONAL MATERIALS FOR CHAPTER 4 PROJECT 3

B.1 MODEL PARAMETERS

Variable N_{user} defines how many users will enter the system and play the game. It also equals to the simulation time step since only one user enters the system at a time. Variable $r_{newquestion}$ is how many new questions entering the system at each time step. Variable $N_{initialquestion}$ controls how many new questions exist when the first user arrives. In a long term, this value shall not affect the overall dynamics. Variable L_w determines how many questions to be selected and be prepared for answering. Early work of page to page surfing and clicking by Huberman *et al.* (1998) gave the probability distribution of number of links. The most portion of frequency of clicks are below five.

Table B.1: Fixed Parameters for Simulation

Fixed Parameter	Value
N_{user} Total number of users	150
$r_{newquestion}$ New question generating rate per user	1
$N_{initialquestion}$ Initial number of questions in the game	10
L_w Window size of questions selection	5
$Prob_{bin}$ Probabilities of number of question bins [1,2,3,4]	[0.221,0.2927,0.2608,0.2256]Li *et al.* (2012)
$Prob_{depth}$ Probabilities of question sequential computing difficulty [2,3,4,5,6,7]	[0.125,0.25,0.3125,0.125,0.125,0.0625] Liu *et al.* (2013)
N_{sim} Number of iterations per simulation configuration	2000
$\beta_1, \beta_2, \beta_3$ Regression coefficients for value computation	[0.83,0.61,0.96] Liu *et al.* (2013)
p_B Probabilities of being in answering preference ["evenly", "until one answer", "ROA"]	[0.3,0.3,0.4]

Variables $Prob_{bin}$ give probabilities of number of bin for 1,2,3 and 4 based on Li *et al.* (2012). Variables $Prob_{depth}$ give probabilities of question sequential computing difficulty or the depth for 2,3,4,5,6 and 7 based on Liu *et al.* (2013). Variable N_{sim} defines the number of iterations during one simulation. All the results are the mean values across all iterations. Parameters β_1, β_2, β_3 are discussed in 3.2.4 and are regression coefficients for value computation. Variable p_B gives the proportion of three categories of attribute B.

B.2 INDEPENDENT STUDY OF RANDOM QUESTION SELECTION

The purpose of this section is to show that the random question selection strategy has similar effect and results as the group with half-half mixing proportion of strategy between selecting easy and difficult questions. We demonstrate variant on the proportion of users who are doing random question selection brings trivial change compared to a population with half-half mixing question selection strategy between selecting easy and selecting difficult questions. Thus the strategy of random question selection can be approximated by a single group of difficulty-biased with half in preference of selecting easy questions and the other half in preference of selecting difficult questions.

In here, again we use the notations of variables X and Y in the experiment of comparison which is different from those in the main paper. Let X be the proportion

of the users randomly selecting questions in the whole population. Let Y be the proportion of users in group of difficulty-biased. Further Y_1 and Y_2 be the proportion of users among all users in favor of easy questions and difficulty questions respectively in the group of difficulty-biased. Therefore,

$$X + Y = 1$$
$$Y = Y_1 + Y_2$$
$$Y_1 = Y_2 = Y/2$$

We vary X from 0 to 1 and take 9 discrete points in between with equal space. Then $X \in [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]$ and $Y \in [1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0]$. When $X = 0$ it means there is no group of random question selection and 100% users are coming from the group of difficulty-biased. When $X = 1$ it means there is no group of difficulty-biased and 100% users are coming from the group of random question selection. The mean of questions answered percentage, average quality and average value across all questions on the community among 2,000 simulations are shown in figure B.1.

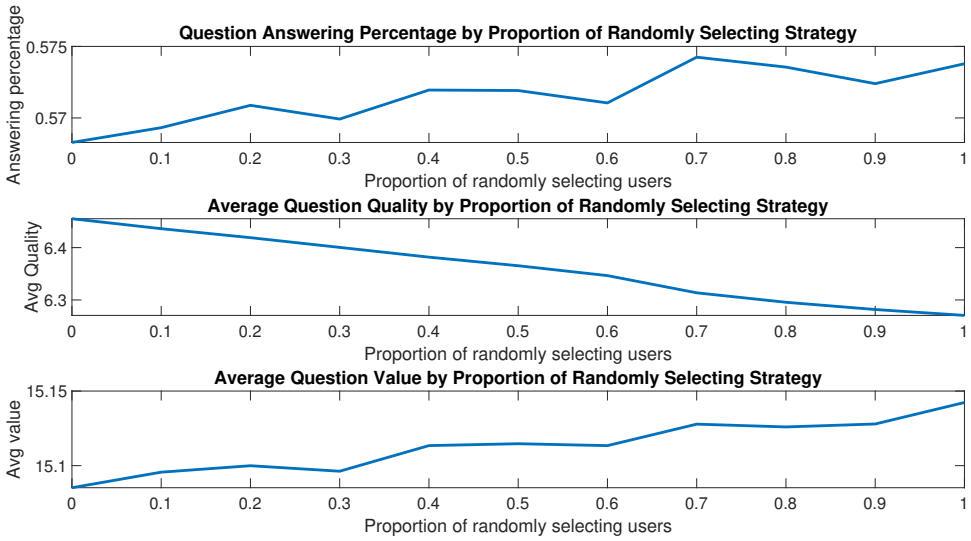

Figure B.1: Comparison of Variant on Random Question Selection (Bars Show One Standard Deviation Ranges)

We see very trivial change from extreme case of 100% users from group of difficulty-biased($X = 0$) to another extreme case of 100% users from group of random question selection ($X = 1$) in the figure B.1. Next we are looking closely into ball distribution among questions between two extreme cases $X = 0$ and $X = 1$.

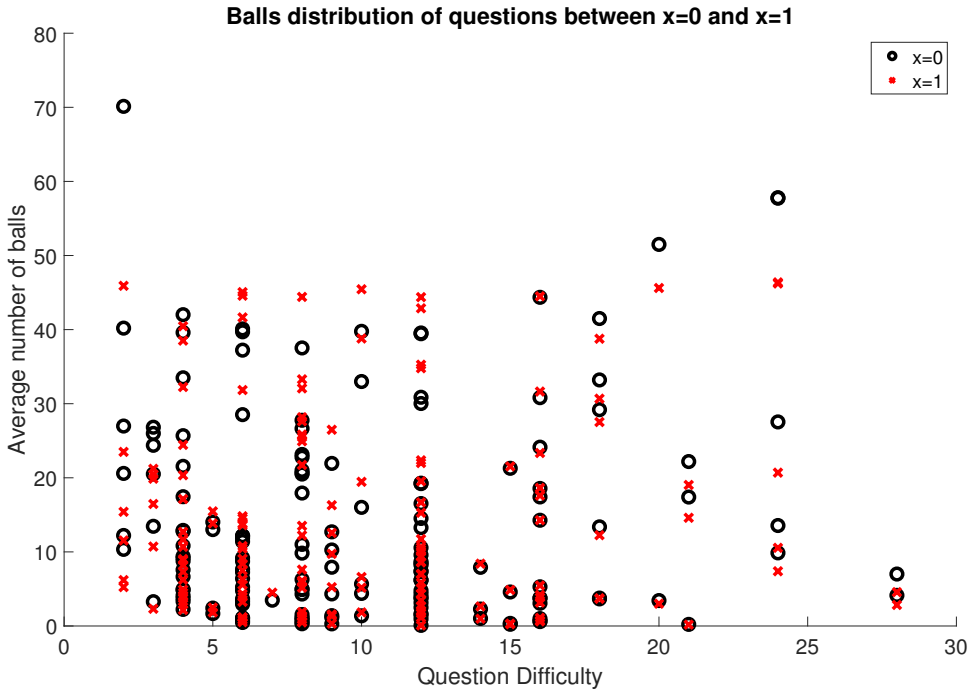

Figure B.2: Comparison of Extreme Cases: All Randomly Selecting and All Difficulty-biased with Half and Half Mixing

By taking mean on number of balls in each question basket among 2,000 simulation runs we show in the figure B.2 there is no significant difference in ball distribution between two extreme case of $x = 0$ and $x = 1$. Each circle and cross in the figure B.2 represent one mean value for one question from two different settings. The case of $x = 0$ has no user from group of random question selection and half of population is in favor of easy questions and the other half in favor of difficult questions. The case of $x = 1$ has users who only randomly select questions to answer.

B.3 SIMULATION RESULT FOR QUESTION SELECTION STRATEGY

B.3.1 BASELINE SCENARIO

We look at the percentage of questions answered on the community for baseline scenario which is half population of difficulty-biased and the other half of popularity-biased. As shown in figure B.3, with a circle we mark the maximum value at each row. For instance at the first row if all users in the group of popularity-biased are in favor of new questions only, then the best outcome is having 80% of users from group of difficulty-biased preferring easy questions and the rest of 20% preferring difficult questions. We perform One-sample Kolmogorov-Smirnov test to examine normality of all 2,000 simulation results and make sure they are normally distributed. Then along each row we run two samples t test for data with maximal sample mean (in circle) and others to draw significant difference thresholds (between lines) with p value 0.05. Every box of mixture with p value less than 0.05 shall be included within the

lines. Every box of mixture outside the lines are considered to be significant lower than optimal value. The area with values between lines forms a optimal path along y axis.

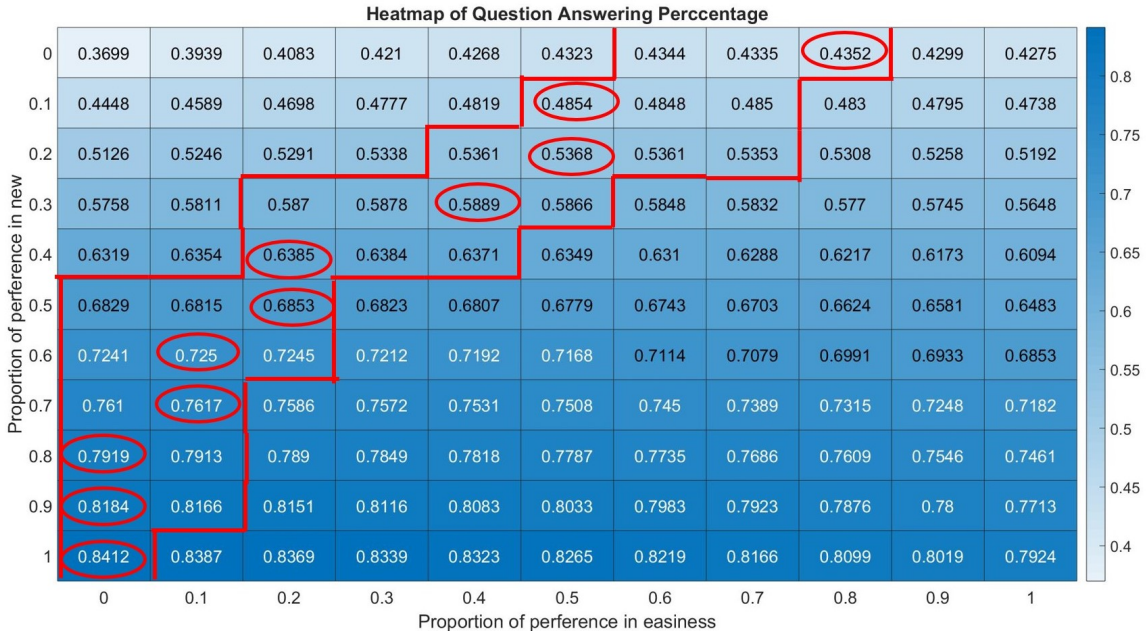

Figure B.3: Proportion Grid of Question Answering Percentage for Baseline Scenario (50% of Popularity-biased and 50% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

Clearly the maximum value on the heatmap B.3 is 0.8412 at the left bottom corner meaning when half of users only selecting new questions and half of users only selecting difficult questions, the overall percentage of questions answered reaches the maximum point 84.12%. Along the circles and between the lines we see an optimal mixture path from top right to bottom left. More users are selecting new questions to answer, higher the percentage of questions answered is. The optimal proportion of users in preference of easiness decreases with increasing number of users in preference of new questions because easy questions are more likely to be addressed already once they are introduced into the system when more users are selecting new questions to answer. We will see similar trend of moving toward bottom left across all three scenarios.

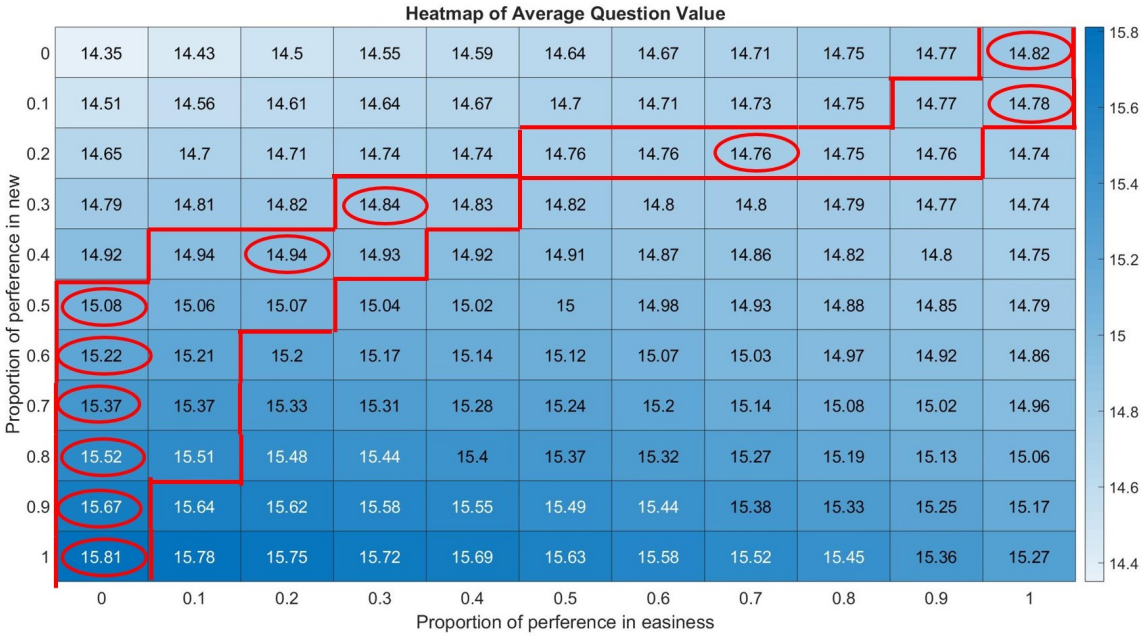

Figure B.4: Proportion Grid of Average Value for Baseline Scenario (50% of Popularity-biased and 50% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

For the average value in the figure B.4, optimal mixture path is toward bottom left with a very steep shifting while there is a small portion of users in favor of new questions (y from 0 to 0.3 in the graph B.4). It seems that there are no need of users in favor of easy questions when there are enough people in favor of new questions. It is similar result as trend of question answering percentage.

Next we present the heatmap for average difficulty of answered questions in figure B.5. Obviously more users selecting difficult questions to answer, higher the average difficulty will be for all answered questions. Heatmap shows one dimensional dynamics as heatmap for average quality B.6. In order to achieve the maximum of average difficulty all in the groups of difficulty-biased shall choose only difficult questions disregarding the mixing proportion in the group of popularity-biased.

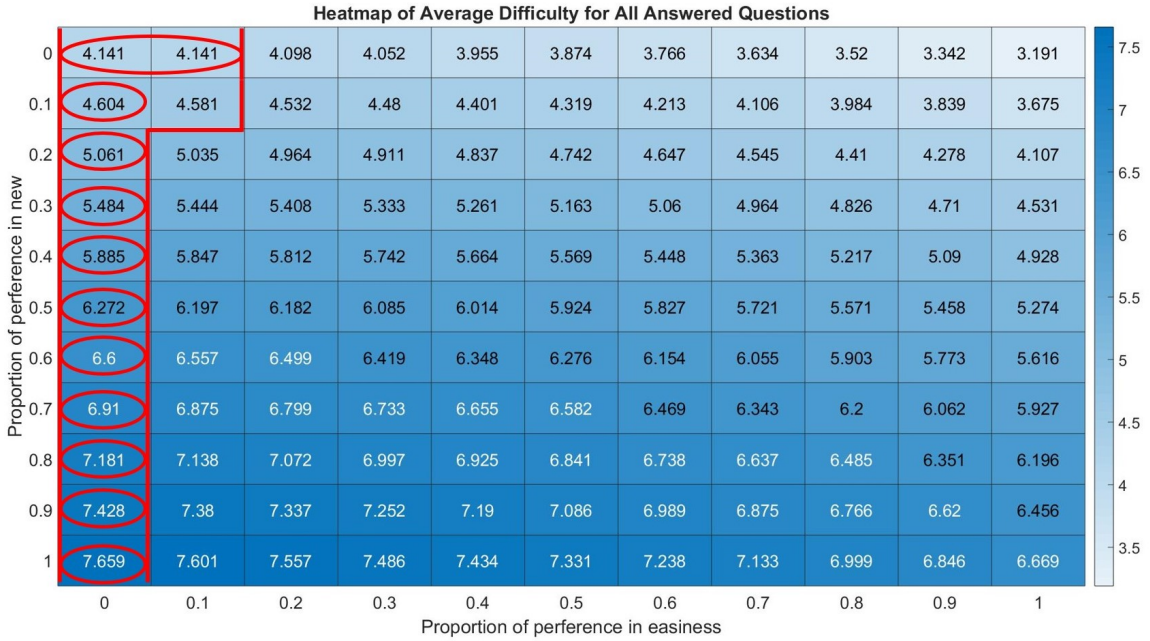

Figure B.5: Proportion Grid of Average Difficulty for User Groups of Popularity-biased and Difficulty-biased with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

At last in figure B.6 in term of average quality of answered questions the maximum value is at the top right corner. In order to have a higher average quality, questions need to accumulate answers. Depth of answers per question scarifies board coverage of questions for answering so best mixing proportion is to have the most people selecting popular questions and easy questions. Such pattern is the same across three scenarios.

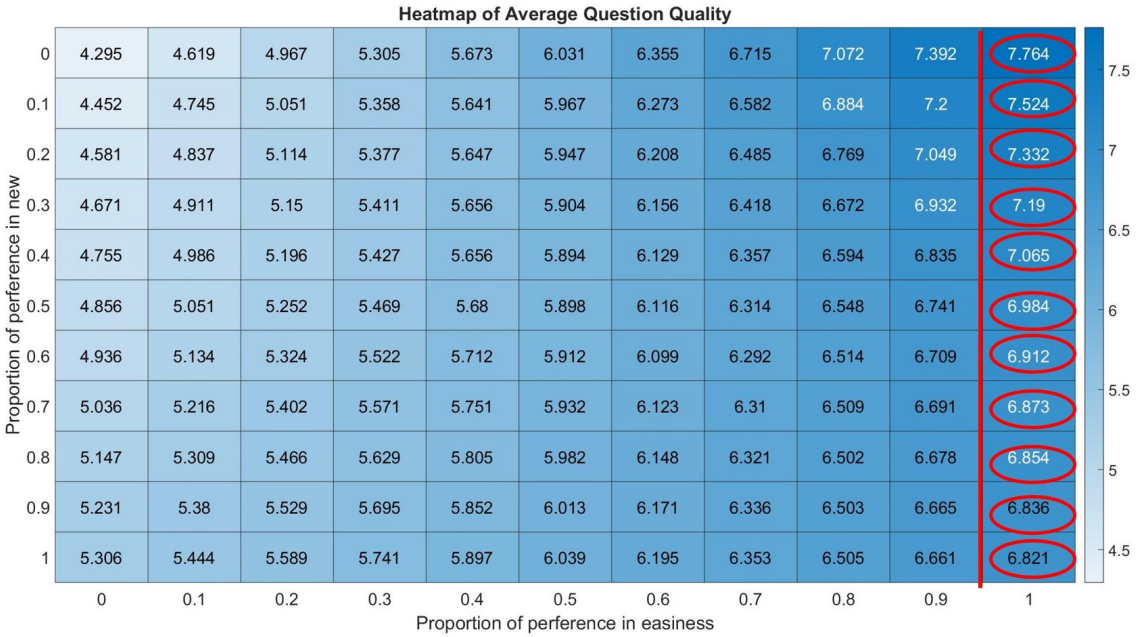

Figure B.6: Proportion Grid of Average Quality for Baseline Scenario (50% of Popularity-biased and 50% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

We compare the results to difficulty dominated scenario and popularity dominated scenario. Moving direction of optimal mixture paths for both are unchanged as we analyzed above for the baseline scenario. The difference is the optimal mixture paths for average percentage of questions answered and average value of questions answered are steeper and sharper changing from one end to another for difficulty dominated scenario with more people in the group of difficulty-biased. For the popularity dominated scenario the change is more gradual in the heatmap of question answering percentage and average value.

B.3.2 POPULARITY DOMINATED SCENARIO

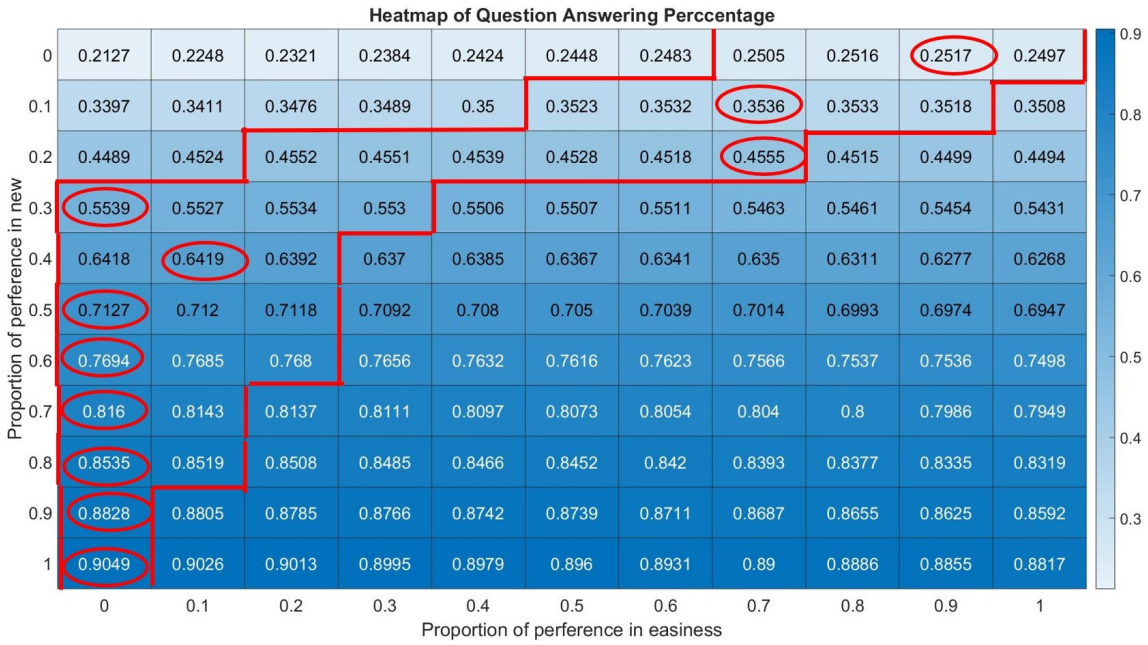

Figure B.7: Proportion Grid of Question Answering Percentage for Popularity Dominated Scenario with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

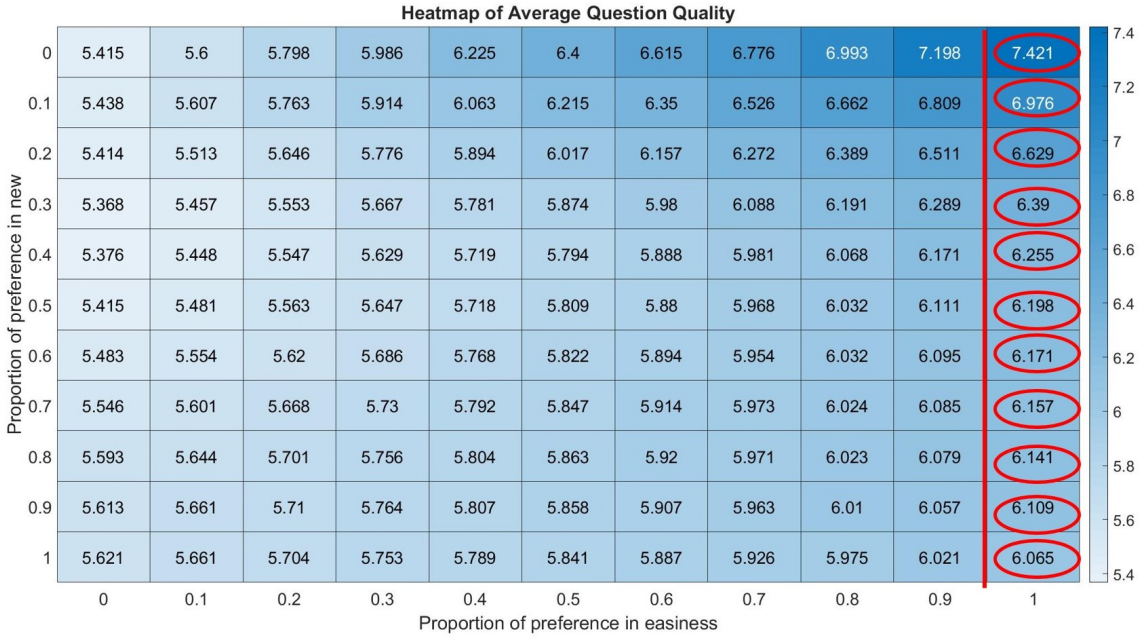

Figure B.8: Proportion Grid of Average Quality for Popularity Dominated Scenario (80% of Popularity-biased and 20% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

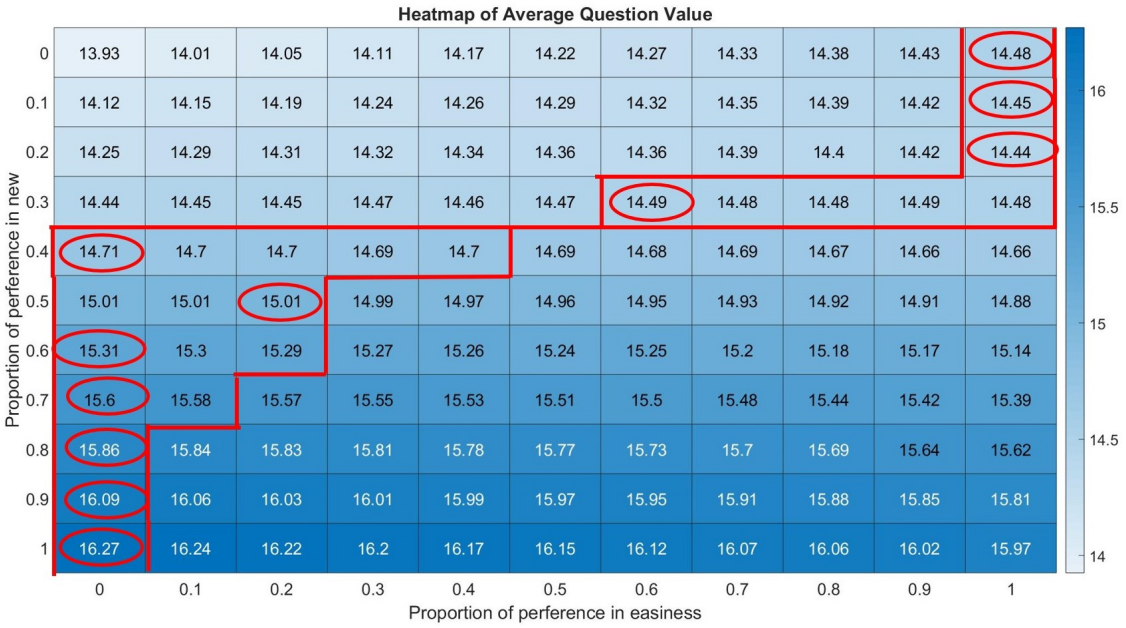

Figure B.9: Proportion Grid of Average Value for Popularity Dominated Scenario(80% of Popularity-biased and 20% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

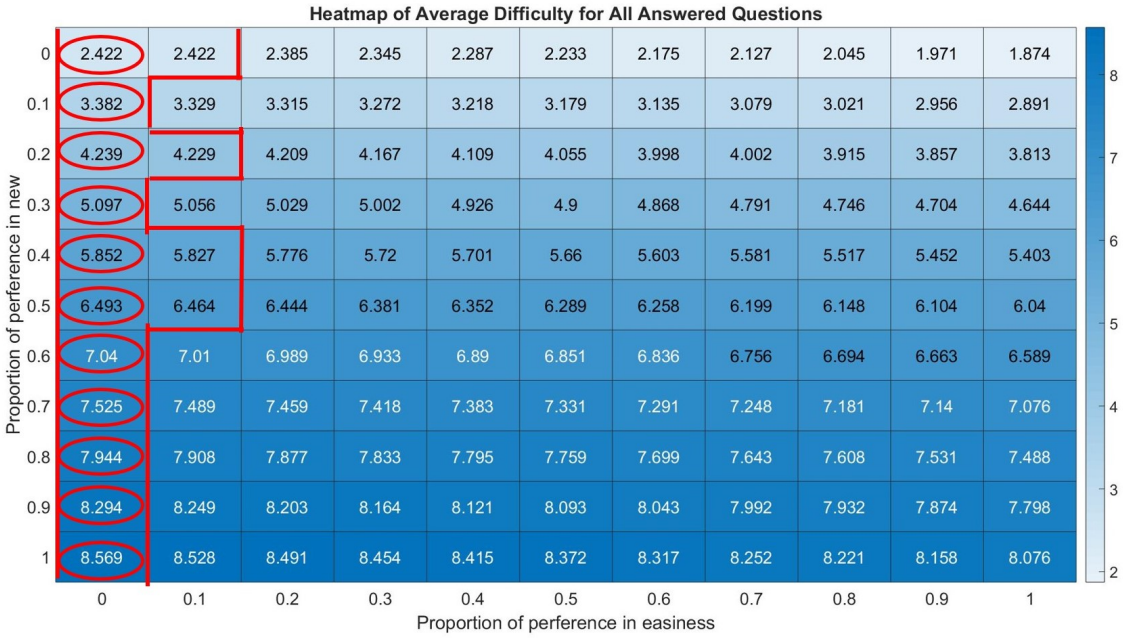

Figure B.10: Proportion Grid of Average Difficulty for Popularity Dominated Scenario(80% of Popularity-biased and 20% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

B.3.3 DIFFICULTY DOMINATED SCENARIO

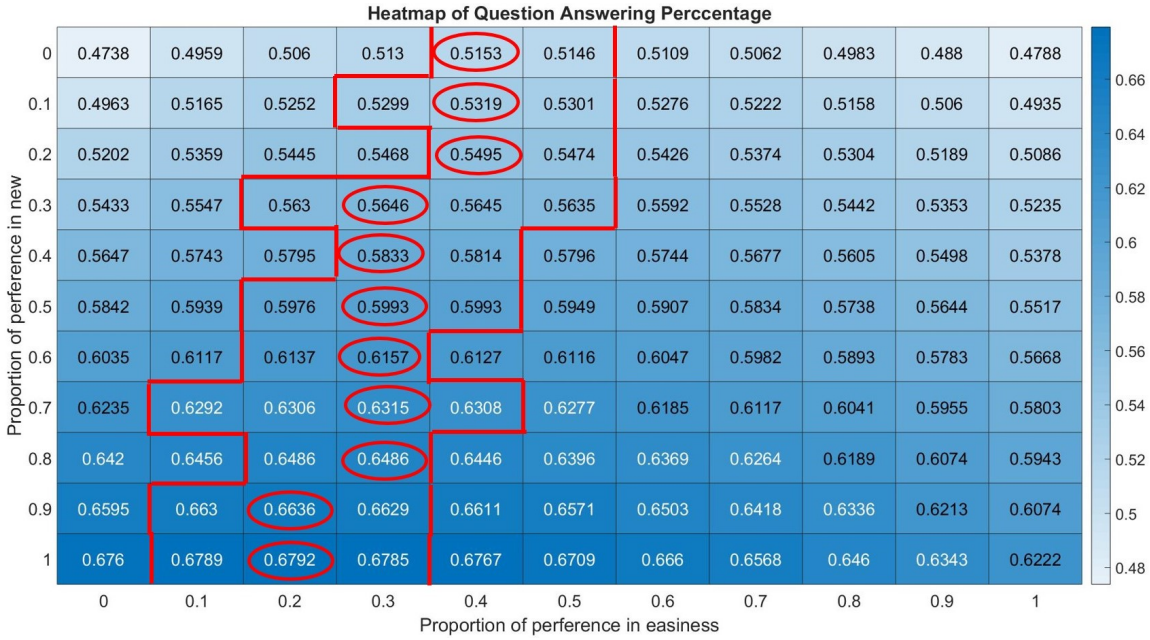

Figure B.11: Proportion Grid of Question Answering Percentage for Difficulty Dominated Scenario (20% of Popularity-biased and 80% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

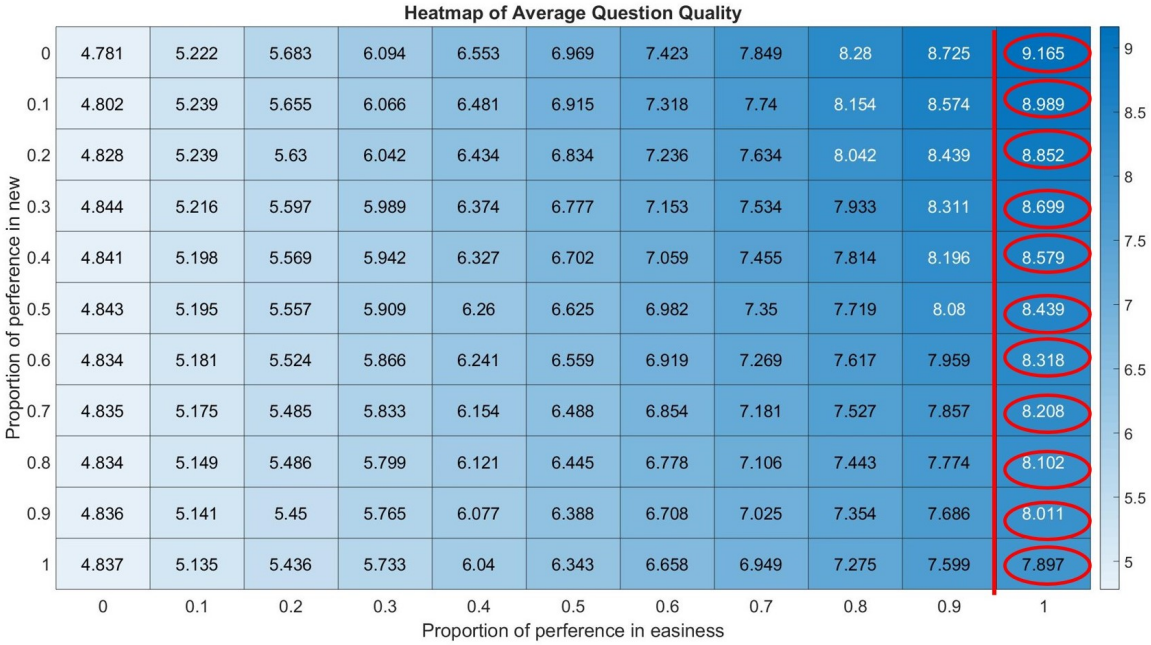

Figure B.12: Proportion Grid of Average Quality for Difficulty Dominated Scenario (20% of Popularity-biased and 80% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

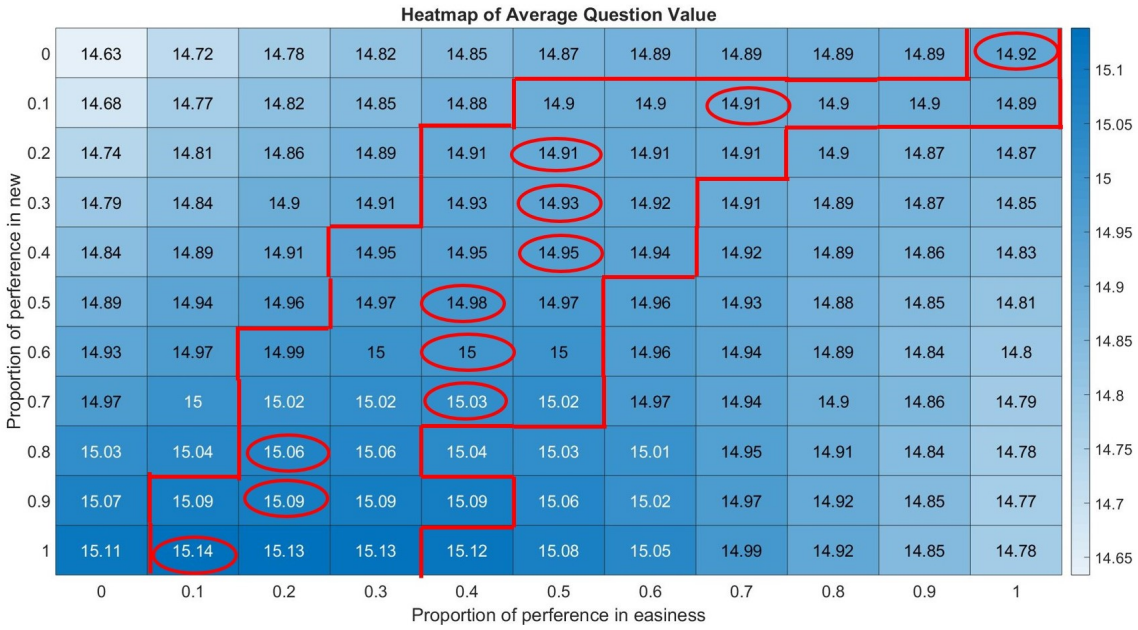

Figure B.13: Proportion Grid of Average Value for Difficulty Dominated Scenario (20% of Popularity-biased and 80% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

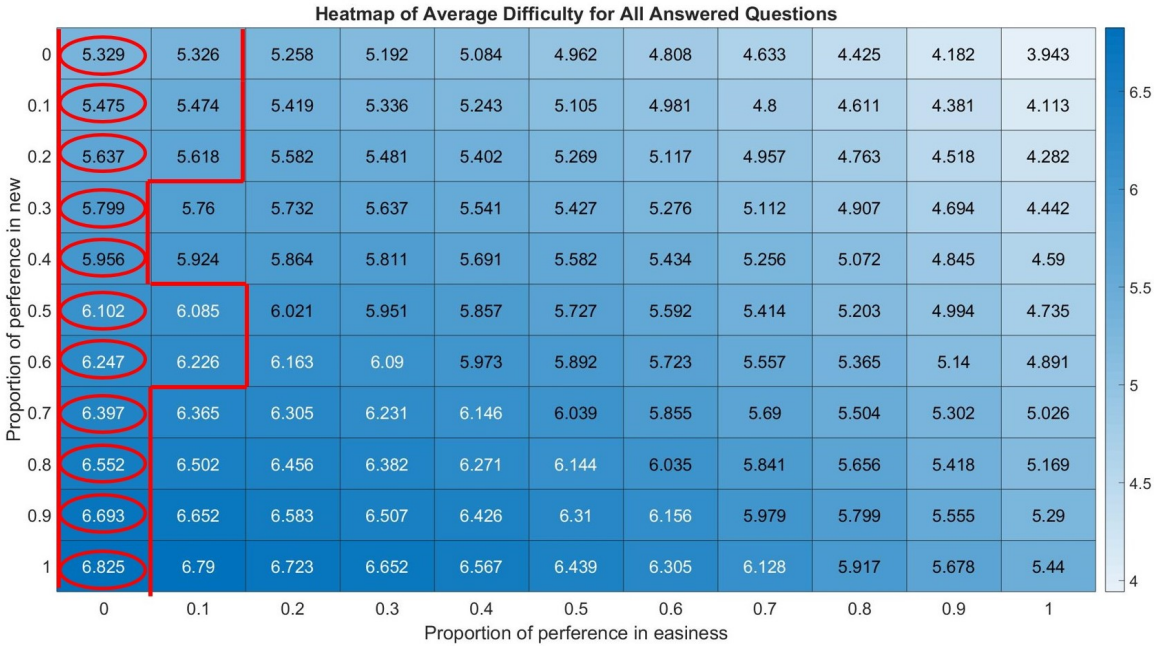

Figure B.14: Proportion Grid of Average Difficulty for Difficulty Dominated Scenario (20% of Popularity-biased and 80% of Difficulty-biased) with Red Circles for the Maximum per Row and Red Lines for the Statistic Insignificant Bond between Two Samples ($p > 0.05$)

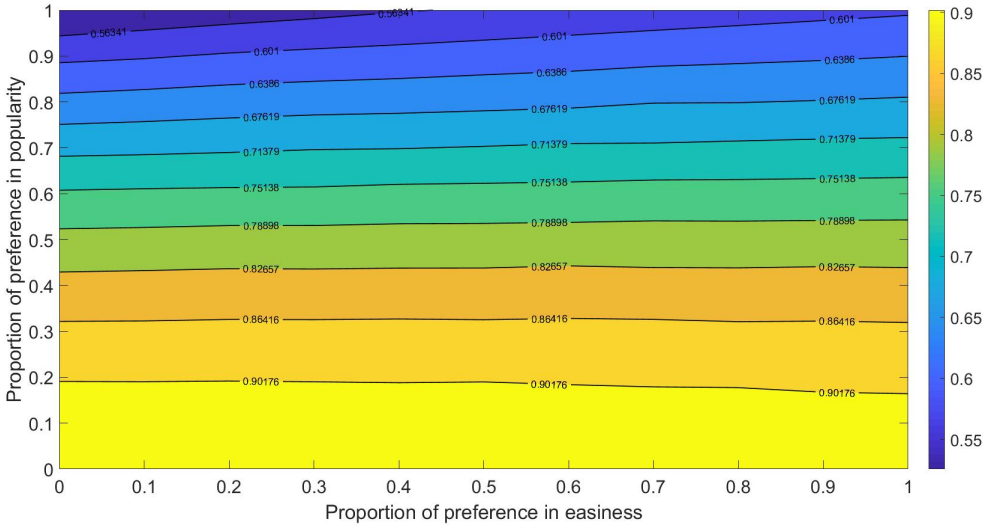

Figure B.15: Contour Map of Equally Weighted Objective Score for Popularity Dominated Scenario