The Size Conundrum: Why Online Knowledge Markets Can Fail at Scale

Himel Dev
University of Illinois (UIUC)
hdev3@illinois.edu

Chase Geigle
University of Illinois (UIUC)
geigle1@illinois.edu

Qingtao Hu
University of Illinois (UIUC)
qhu13@illinois.edu

Jiahui Zheng
Peking University
alicehz@pku.edu.cn

Hari Sundaram
University of Illinois (UIUC)
hs1@illinois.edu

ABSTRACT

In this paper, we interpret the community question answering websites on the StackExchange platform as knowledge markets, and analyze how and why these markets can fail at scale. A knowledge market framing allows site operators to reason about market failures, and to design policies to prevent them. Our goal is to provide insights on large-scale knowledge market failures through an interpretable model. We explore a set of interpretable economic production models on a large empirical dataset to analyze the dynamics of content generation in knowledge markets. Amongst these, the Cobb-Douglas model best explains empirical data and provides an intuitive explanation for content generation through the concepts of elasticity and diminishing returns. Content generation depends on user participation and also on how specific types of content (e.g. answers) depends on other types (e.g. questions). We show that these factors of content generation have constant elasticity—a percentage increase in any of the inputs leads to a constant percentage increase in the output. Furthermore, markets exhibit diminishing returns—the marginal output decreases as the input is incrementally increased. Knowledge markets also vary on their returns to scale—the increase in output resulting from a proportionate increase in all inputs. Importantly, many knowledge markets exhibit diseconomies of scale—measures of market health (e.g., the percentage of questions with an accepted answer) decrease as a function of the number of participants. The implications of our work are two-fold: site operators ought to design incentives as a function of system size (number of participants); the market lens should shed insight into complex dependencies amongst different content types and participant actions in general social networks.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Social networks**;

KEYWORDS

community question answering; knowledge market; content generation; diseconomies of scale

1 INTRODUCTION

In this paper, we analyze a large group of community question answering (CQA) websites on StackExchange network through the Economic lens of a market. Framing StackExchange sites as knowledge markets has intuitive appeal: in a hypothetical knowledge market, if no one wants to answer questions, but only ask, or conversely, there are individuals who want to only answer but not ask questions, the "market" will collapse. What, then, is the required relationship among actions (say between questions and answers) in such a knowledge market for us to deem it healthy? Are larger markets with more participants healthier since there will be more people to ask and answer questions?

Studying CQA websites through an economic lens allows site operators to reason about whether they should grow the user base. Since most of the popular CQA websites (e.g. Quora, StackExchange) do not charge participants, but instead depend on site advertisements for revenue, there is a natural temptation for operators of these sites to grow the user base so that there is increase in revenue. As we show in this paper, for most StackExchange sites, growth in the user base is counter-productive in the sense that they turn unhealthy—specifically, more questions remain unanswered.

Explaining the macroscopic behavior of knowledge markets is important, yet challenging. One can regress some variable of interest (say number of questions) on variables including number of users, time spent in the website among others. However, explaining why the regression curve looks the way it does is hard. As we show in this work, using an economic lens of a market allows us to model dependencies between number of participants and the amount of content, and to predict the production of content.

Our main contribution is to model CQA websites as knowledge markets, and to provide insight on the relationship between size and health of these markets. To this end, we develop models to capture content generation dynamics in knowledge markets. We analyze a set of basis functions (the functional form of how an input contributes to output) and interaction mechanisms (how the inputs interact with each other), and identify the optimal *power basis* function and the *interactive essential* interaction form using a prediction task on the outputs (questions, answers, and comments).

This form is the well-known Cobb-Douglas form that connects production inputs with output. Using the best model fits for each StackExchange site, we show that the Cobb-Douglas model predicts the production of content with high accuracy.

The Cobb-Douglas function provides an intuitive explanation for content generation in StackExchange markets. It demonstrates that, in StackExchange markets, (1) factors such as user participation and content dependency have *constant elasticity*—percentage increase in any of these inputs will have constant percentage increase in output; (2) in many markets, factors exhibit *diminishing returns*—decrease in the marginal (incremental) output (e.g., answer production) as an input (e.g. number of people who answer) is incrementally increased, keeping the other inputs constant; (3) markets vary according to their *returns to scale*—the increase in output resulting from a proportionate increase in all inputs; and (4) many markets exhibit *diseconomies of scale*—measures of health decrease as a function of overall system size (number of participants).

There are two reasons why we see diminishing returns in the StackExchange markets. First, the total activity of participants for any StackExchange market unsurprisingly follows a power-law pattern. What is interesting is that the power-law exponent falls with increase in size for most markets, implying that new users do not participate in the same manner as earlier users. Second, we can identify a stable core of users who actively participate for long periods of time, contributing to the market health.

Finally, we show diseconomies of scale through experiments on system size, analysis of health metrics, and user exchangeability. For most StackExchange markets, we see that as system size grows, the ratio of answers to questions falls below a critical point, when some questions go unanswered. Furthermore, using health metrics of the number of questions with an accepted answer, and the number of questions with at least one answer, we observe that most StackExchange markets decline in health with increase in size. Finally, we compare the top contributors with the bottom contributors to see if they are "exchangeable." Most StackExchange markets are not exchangeable in the sense the contributions of the top and the bottom contributors are qualitatively different and differ in absolute terms. These experiments on diseconomies of scale are consistent with the insight from Cobb-Douglas model of production that predicts diminishing returns.

2 RELATED WORK

Our work draws from, and improves upon, several research threads.

Sustainability. Srba and Bielikova [35] conducted a case study on why StackOverflow, the largest and oldest of the sites in StackExchange network, is failing. They shed some insights into knowledge market failure such as novice and negligent users generating low quality content perpetuating the decline of the market. However, they do not provide a systematic way to understand and prevent failures in these markets. Wu et al. [41] introduced a framework for understanding the user strategies in a knowledge market—revealing the importance of diverse user strategies for sustainable markets. In this paper, we present an alternative model that provides many interesting insights including knowledge market sustainability.

Activity Dynamics. Walk et al. [39] modeled user-level activity dynamics in StackExchange using two factors: intrinsic activity

decay, and positive peer influence. However, the model proposed there does not reveal the collective platform dynamics, and the eventual success or failure of a platform. Abufouda [1] developed two models for predicting the interaction decay of community members in online social communities. Similar to Walk et al. [39], these models accommodate user-level dynamics, whereas we concentrate on the collective platform dynamics. Wu and Zhang [42] proposed a discrete generalized beta distribution (DGBD) model that reveals several insights into the collective platform dynamics, notably the concept of a size-dependent distribution. In this paper, we improve upon the concept of a size-dependent distribution.

Economic Perspective. Kumar et al. [24] proposed an economic view of CQA platforms, where they concentrated on the growth of two types of users in a market setting: users who provide questions, and users who provide answers. In this paper, we concentrate on a subsequent problem—the "relation" between user growth and content generation in a knowledge market. Butler [9] proposed a resource-based theory of sustainable social structures. While they treat members as resources, like we do, our model differs in that it concentrates on a market setting, instead of a network setting, and takes the complex content dependency of the platform into consideration. Furthermore, our model provides a systematic way to understand successes and failures of knowledge markets, which none of these models provide.

Scale Study. Lin et al. [27] examined Reddit communities to characterize the effect of user growth in voting patterns, linguistic patterns, and community network patterns. Their study reveals that these patterns do not change much after a massive growth in the size of the user community. Yla et al. [44] investigated the effects of crowd size on solution quality in StackExchange communities. Their study uncovers three distinct levels of group size in the crowd that affect solution quality: topic audience size, question audience size, and number of contributors. In this paper, we examine the consequence of scale on knowledge markets from a different perspective by using a set of health metrics.

Stability. Successes and failures of platforms have been studied from the perspective of user retention and stability [12, 16, 22, 30]. Notably, Patil et al. [30] studied the dynamics of group stability based on the average increase or decrease in member growth. Our paper examines stability in a different manner—namely, by considering the relative exchangeability of users as a function of scale.

User Growth. Successes and failures of user communities have also been widely studied from the perspective of user growth [6, 21, 25, 32, 45]. Kairam et al. [21] examined diffusion and non-diffusion growth to design models that predict the longevity of social groups. Ribeiro [32] proposed a daily active user prediction model which classifies membership based websites as sustainable and unsustainable. While this perspective is important, we argue that studying the successes and failures of communities based on content production can perhaps be more meaningful [23, 47, 48].

Modeling CQA Websites. There is a rich body of work that extensively analyzed CQA websites [3, 5, 10, 34, 40], along with user behavior [4, 19, 31, 38, 46], roles [15, 26], and content generation [7, 14, 43]. Notably, Yang et al. [43] noted the *scalability problem* of CQA—namely, the volume of questions eventually subsumes the capacity of the answerers within the community. Understanding and modeling this phenomenon is one of the goals of this paper.

3 PROBLEM FORMULATION

The goal of this paper is to develop a model for content generation in knowledge markets. Content is integral to the success and failure of a knowledge market. Therefore, we aim to better understand the content generation dynamics.

A model for content dynamics should have the following properties: macro-scale, explanatory, predictive, minimalistic, comprehensive.

Macro-scale: The model should capture content generation dynamics via aggregate measures. Aggregate measures help us understand the collective market by summarizing a complex array of information about individuals, which is especially important for policy-making.

Explanatory: The model should be insightful about the behavior of a knowledge market. Understanding market behavior is a crucial first step in designing policies to maintain a resilient, sustainable market.

Predictive: The model should allow us to make predictions about future content generation and resultant success or failure. These market predictions are integral to the prevention and mitigation of market failures.

Minimalistic: The model should have as few parameters as necessary, and still closely reflect the observed reality.

Comprehensive: The model should encompass content generation dynamics for different content types (e.g., question, answer, comment) in varieties of knowledge markets. This is important for developing a systematic way to understand the successes and failures of knowledge markets.

In remaining sections we propose models that meet the aforementioned requirements, and show that our best-fit model accurately reflects the content generation dynamics and resultant successes and failures of real-world knowledge markets.

4 MODELING KNOWLEDGE MARKETS

In this section, we introduce economic production models to capture content generation dynamics in real-world knowledge markets. We first draw an analogy between economic production and content generation, and report the content generation factors in knowledge markets (Section 4.1). Then, we concentrate on the knowledge markets in StackExchange—presenting production models for different content types (Section 4.2).

4.1 Preliminaries

Economic production mechanisms well describe content generation in knowledge markets. In economics, *production* is defined as the process by which human labor is applied, usually with the help of tools and other forms of capital, to produce useful goods or services—the *output* [36]. We assert that participants of a knowledge market function as labor to generate content such as questions and answers. Analogous to economic output, content contributes to participant utility.

Motivated by the production analogy, we design macroeconomic production models to capture content generation dynamics in knowledge markets. In these models, instead of directly modeling content generation as a dynamic process (function of time), we model it in terms of associated factors which are dynamic.

There are two key factors that affect content generation in knowledge markets, namely user participation and content dependency. User participation is the most important factor in deciding the quantity of generated content. The participation of more users induce more questions, answers, and other contents in a knowledge market. Content dependency also affects the quantity of generated content for different types. Content dependency refers to the dependency of one type of content (e.g., answers) on other type of content (e.g., questions). In absence of questions, there will be no answers in a knowledge market, even in the presence of many potential participants who are willing to answer.

4.2 Modeling StackExchange

StackExchange is a network of community question answering websites where each site is based on a focused topic. Each user of the StackExchange network participates in one or more of these sites based on their interests. StackExchange sites are free knowledge markets where participants generate content for non-monetary reputation-based incentives. These markets are diverse, varying in theme (subject matter), size (number of users and amount of activity), and age (number of days in existence).

We design production models for three primary content types in StackExchange: questions (the root content), answers (which nest below questions), and comments (which can nest either beneath questions or answers). Based on the content dependency and user roles in content generation, we propose the following relationships for question, answer and comment generation in StackExchange (see Table 1 for notation).

Table 1: Notation used in the model

Symbol	Definition
$U_q(t)$	# of users who asked questions at time t
$U_a(t)$	# of users who answered questions at time t
$U_c(t)$	# of users who made comments at time t
$N_q(t)$	# of active questions at time t
$N_a(t)$	# of answers to active questions at time t
$N_c^q(t)$	# of comments to active questions at time t
$N_c^a(t)$	# of comments to active answers at time t
$N_c(t)$	# of comments to active questions/answers at time t
f_x	The functional relationship for content type x

There is a single factor in generating N_q questions: the number of users U_q who ask questions (askers).

$$N_q = f_q(U_q)$$

There are two key factors in generating N_a answers: the number of questions N_q, and the number of users U_a who answer questions (answerers).

$$N_a = f_a(N_q, U_a)$$

There are two types of comments: comments on questions, and comments on answers. Accordingly, there are three key factors in generating N_c comments: the number of questions N_q, the number of answers N_a, and the number of users U_c who make comments (commenters).

$$N_c^q = f_{c^q}(N_q, U_c)$$

$$N_c^a = f_{c^a}(N_a, U_c)$$

$$N_c = N_c^q + N_c^a$$

The aforementioned relationships imply that the amount of generated content of each type depends on the function describing its factor dependent growth, and the availability of factor(s). These relationships make three assumptions. First, different content types interact only through their use of factors. Second, the functional relationships depend on the consumption or usage of each factor. Third, the functional relationships depend on the interaction among the factors—how the factors of a particular content type interact.

Now, we transform the functional relationships into production models by first choosing a basis function to capture how a content type consumes its factor(s), and then choosing an interaction type to capture the interaction among factors.

Basis Function. We use a basis function to capture the effect of a given factor on a particular content type. While there is a variety of basis functions available for regression, we consider three basis functions widely used in economics and growth modeling [13]: power— $g(x) = ax^\lambda$; exponential— $g(x) = ab^x$; and sigmoid— $g(x) = \frac{L}{1+e^{k(x-x_0)}}$.

Interaction among the Factors. We use an aggregate function to capture the interaction among multiple factors of a given content type. Specifically, we consider the pairwise interaction functions listed in Table 2.

We combine a basis function and an interaction type to design production models for different content types. For example, answer generation can be modeled using power basis and essential interaction as $N_a = min(a_1 N_q^{\lambda_1}, a_2 U_a^{\lambda_2})$. We consider twelve such possible models (combination of three basis and four interaction type) for answer and comment generation in StackExchange.

User Role Distribution. A fundamental assumption of our model is the awareness of user roles (e.g., asker, answerer, and commenter) and their distribution (e.g., how many users are askers?). We empirically observe that all StackExchange markets have a stable distribution of user roles. In fact, given the number of users, we can accurately predict the number of participants for each role.

We apply linear regression to determine the number of participants U_x of a particular role $x \in \{q, a, c\}$ from the number of users U in a StackExchange market. For each market, we compute three distinct coefficients of determination, R^2, for predicting three roles (asker, answerer, and commenter) using linear regression. In Figure 1 we show the distribution of R^2 for regressing user roles across 156 StackExchange markets. We use letter value plots[1] to present these distributions—showing precise estimates of their tail behavior. We observe that, in most markets, the R^2 values are close to 1. Further, the tail capturing low R^2 values consists of markets with a relatively small number of monthly users.

Number of Users. The number of users is the only free input to our content generation models; the remaining inputs are functions of the number of users. In all these models, the growth or decline of number of users is exogenous—determined outside the model, by non-economic forces.

---

[1]The letter-value plot display information about the distribution of a variable [20]. It conveys precise estimates of tail behavior using letter values; boxplots lack such precise estimation.

Table 2: Pairwise interaction between factors with contour

<u>Essential:</u> Essential factors are both required for content generation, with zero marginal return for a single factor. For a pair of essential factors, content generation is determined by the more limiting factor: $z = min(y_1, y_2)$ [37]. This is known as Liebig's law of the minimum.

<u>Interactive Essential:</u> In interactive essential interaction, we get diminishing return (instead of zero return) for a single factor: $z = y_1 y_2$ [37]. If factors are consumed using power basis function, i.e., $y_i = ax_i^\lambda$, it captures Cobb-Douglas production function.

<u>Antagonistic:</u> For antagonistic factors, content generation is determined solely by the availability of the factor which yields the largest return: $z = max(y_1, y_2)$ [37]. This interaction implies that the production process has maximum possible efficiency.

<u>Substitutable:</u> Factors that can each support production on their own are substitutable relative to each other: $z = w_1 y_1 + w_2 y_2$ [37]. This implies that there exists some equivalence between the two factors. This is analogous to the general additive models.
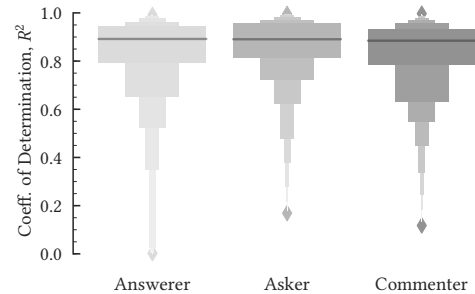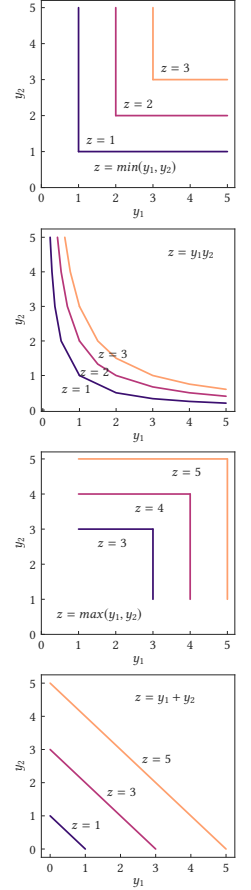
Figure 1: The distribution of coefficient of determination, R^2, for regressing user roles across 156 StackExchange. In most StackExchange, the role distribution is stable—as manifested by the arrangement of R^2 in letter-value plots.

5 DATASET

We collected the latest release (September, 2017) of the StackExchange dataset. This snapshot is a complete archive of all activities in StackExchange sites. There are 169 sites in our collected dataset.

For the purpose of empirical analysis, we only consider the sites that have been active for at least 12 months beyond the *ramp up* period (site created, but few or no activity). There are 156 such sites. The age of these sites vary from 14 months to 111 months, number of users from 1,072 to 547,175, and the number of posts (questions and answers) from 1,600 to 1,985,869. Further, the sites have small overlaps in user base; therefore, we can reasonably argue that the underlying markets are independent.
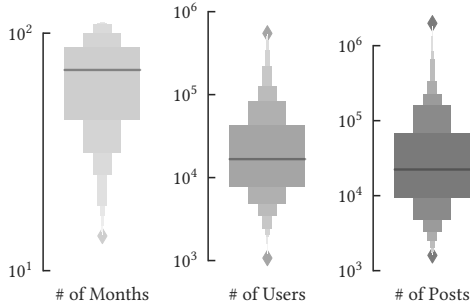
Figure 2: The log-scale distribution of number of months (age), number of users, and number of posts for 156 Stack-Exchange sites.

In Figure 2 we present letter value plots (in log-scale) to show the distribution of number of months (age), number of users, and number of posts for 156 StackExchange sites.

6 EVALUATING OUR PROPOSED MODELS

In this section, we identify optimal models (basis and interaction) based on the accuracy of fitting content generation time series observed in our dataset (Section 6.1), and evaluate the performance of the optimal models in predicting content volume in long run (Section 6.2).

6.1 Model Fitting

We fit each variant of production model (basis and interaction), for each content type, to the observed content generation time series (monthly granularity), in each StackExchange site. Notice that among the different variants of production models, the models using power or exponential basis have a parsimonious set of parameters. For example, answer generation model using power basis function requires only three parameters for interactive essential interaction (See Section 4.2), and four parameters for remaining interaction types. In contrast, answer generation model using sigmoid basis function requires five parameters for interactive essential interaction, and six parameters for remaining interaction types.

Parameter Estimation. We learn the best-fit parameters for capturing the observed content generation time series. We restrict some parameters of our production models to be non-negative, e.g., non-negative exponents in power basis. These restrictions are important because the underlying factors positively affect the output. We use the trust-region reflective algorithm [8] to solve our constrained least square optimization problem. The algorithm is appropriate for solving non-linear least squares problems with constraints.

Evaluation Method. We evaluate fitting accuracy using four metrics: root mean square error (RMSE), normalized root mean square error (NRMSE), explained variance score (EVS), and Akaike information criterion (AIC). Given two series for each content type, the observed series $N(t)$, and the prediction $\hat{N(t)}$ of the series by a model with k parameters, we compute the four metrics as follows:

RMSE $= \sqrt{\frac{1}{T} \sum_{t=1}^{T} (N(t) - \hat{N(t)})^2}$; NRMSE $= \frac{RMSE}{max(N(t)) - min(N(t))}$;

EVS $= 1 - \frac{Var(N(t) - \hat{N(t)})}{Var(N(t))}$; AIC $= T * ln(\frac{1}{T} \sum_{t=1}^{T} (N(t) - \hat{N(t)})^2) + 2k$.

Among the four metrics, RMSE and NRMSE are error metrics (low value implies good fit), AIC is an information theoretic metric to capture the trade-off between model complexity and goodness-of-fit (low value implies good model), and EVS refers to a model's ability to capture variance in data (high value implies good model).

Fitting Results. We compare the fitting accuracy of production models for all StackExchange sites using the four metrics. Each metric is summarized via the mean, across all sites, for each content type. We use content generation time series with monthly granularity as observed data. We found that the models with the exponential and sigmoid basis functions do not fit the data for many StackExchange sites. Accordingly, in Table 3 we only present the results for production models with the power basis and different interaction types. Notice that the models with interactive essential interaction outperform the remaining models for all metrics and content types. We performed paired t-tests to determine if the improvements for interactive essential interaction are statistically significant; the results are positive with $p < 0.01$.

Table 3: The comparison of fitting accuracy of production models (with power basis and different interaction types) for all StackExchange sites. The models with interactive essential interaction outperform the remaining models for all metrics and content types. The improvements for interactive essential interaction are statistically significant, validated via paired t-tests, where $p < 0.01$.

Content	Interaction Type	Avg. RMSE	Avg. NRMSE	Avg. EVS	Avg. AIC
Question	Single Factor	25.74	0.09	0.79	104.47
Answer	Essential	70.31	0.09	0.79	208.82
	I. Essential	**64.62**	**0.08**	**0.83**	**196.39**
	Antagonistic	72.77	0.09	0.78	210.96
	Substitutable	68.90	0.09	0.81	207.61
Comment	Essential	146.64	0.08	0.83	328.25
	I. Essential	**137.23**	**0.08**	**0.85**	**318.24**
	Antagonistic	155.97	0.09	0.82	334.12
	Substitutable	155.43	0.09	0.82	335.10

Thus we use production models with power basis and interactive essential interaction for prediction tasks.

6.2 Forecasting Content Generation

We apply production models with power basis and interactive essential interaction to forecast content volume in long run—one

year ahead in the future. Specifically, we train each model using the content generation data from the first 12 months (beyond the ramp period), and then examine how well the model forecasts content dynamics in the next 12 months. We validate the forecasting capability by examining the overall prediction error (NRMSE).

We compute the prediction NRMSE across all StackExchange sites, and summarize the results using the mean (μ) and variance (σ)— (i) question: $\mu = 0.11$, $\sigma = 0.08$; (ii) answer: $\mu = 0.12$, $\sigma = 0.09$; (iii) comments: $\mu = 0.11$, $\sigma = 0.10$. Notice that our models can forecast future content dynamics with high accuracy. We performed these experiments for different time granularity, e.g., week, month, quarter, and reached a consistent conclusion. We do not report these results for brevity.

7 CHARACTERIZING KNOWLEDGE MARKETS

In this section, we characterize the knowledge markets in StackExchange. We explain the best-fit models and their foundations (Section 7.1), reveal two key distributions that control the markets (Section 7.2), and uncover the stable core that maintains market equilibrium (Section 7.3).

7.1 Model Interpretation

First, we explain the best-fit models found in Section 6.1. We observe that content generation in StackExchange markets are best modeled through the combination of power basis and interactive essential interaction. In addition, we found that the best-fit exponents (λ parameter in basis $g(x) = ax^\lambda$, where x is a factor) of these models lie between 0 and 1 (inclusive), for all factors of all content types, for all StackExchange markets.

A model that uses the power basis (where exponents lie between 0 and 1) and interactive essential interaction is known as the Cobb-Douglas production function [18]. In its most standard form for production of a single output z with two inputs x_1 and x_2, the function is:

$$z = ax_1^{\lambda_1} x_2^{\lambda_2}.$$

Here, the coefficient a represents the *total factor productivity*—the portion of output not explained by the amount of inputs used in production [18]. As such, its level is determined by how efficiently the inputs are utilized in production. The exponents λ_i represent the *output elasticity* of the inputs—the percentage change in output that results from the percentage change in a particular input [18].

The Cobb-Douglas function provides intuitive explanation for content generation in StackExchange markets. In particular, the explanation stands on three phenomena or principles: constant elasticity, diminishing returns, and returns to scale.

Constant Elasticity. In StackExchange markets, factors such as user participation and content dependency have *constant elasticity*—percentage increase in any of these inputs will have constant percentage increase in output [18], as claimed by the corresponding exponents in the model. For example, in ACADEMIA ($N_a = 6.93 N_q^{0.18} U_a^{0.65}$), a 1% increase in number of answerers (U_a) leads to a 0.65% increase in number of answers (N_a).

Diminishing Returns. For a particular factor, when the exponent is less than 1, we observe *diminishing returns*—decrease in

the marginal (incremental) output as an input is incrementally increased, while the other inputs are kept constant [18]. This "law of diminishing returns" has many interesting implications for the StackExchange markets, including the diminishing benefit of having a new participant in a market. For example, in ACADEMIA, if the number of answerers is 100, then the marginal contribution of a new answerer is $c(101^{0.65} - 100^{0.65}) = 0.129c$, where c is a constant; in contrast, if the number of answerers is 110, then the marginal contribution of a new answerer is $c(111^{0.65} - 110^{0.65}) = 0.125c$. Thus, for answer generation in ACADEMIA, including a participant when the number of participants (system size) is 110 is likely to be less beneficial compared to including a participant when the system size is 100.

Returns to scale. The knowledge markets in StackExchange vary in terms of scale efficiency, as manifested by their *returns to scale*—the increase in output resulting from a proportionate increase in all inputs [18]. If a market has high returns to scale, then greater efficiency is obtained as the market moves from small- to large-scale operations. For example, in ACADEMIA, for answer generation, the returns to scale is $0.18 + 0.65 = 0.83 < 1$. The market becomes less efficient as answer generation is expanded, requiring more questions and answerers to increase the number of answers by same amount.

7.2 Two Key Distributions

Next, we discuss two key distributions that control content generation in knowledge markets, namely participant activity and subject POV (perspective). These two distributions induce the three phenomena reported in section 7.1.

Participant Activity. The distribution of participant activities implicitly drives a market's return in terms of user participation, as manifested by the corresponding exponent. For example, in a hypothetical knowledge market where each answerer contributes equally, the answer generation model should be $N_a = A N_q^{\lambda_1} U_a^{1.0}$. In reality, the distribution of participant activities is a size dependent distribution controlled by the number of participants (system size). As the system size increases, most participants contribute to the head of the distribution (few activities), whereas very few join the tail (many activities).
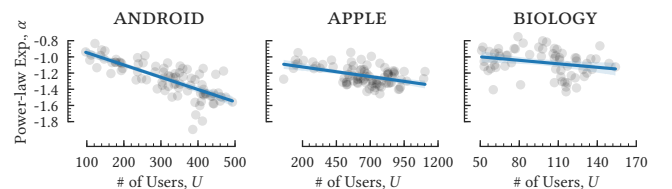
Figure 3: The visibility of size dependent distribution: strong—ANDROID; moderate—APPLE; and weak—BIOLOGY. In most StackExchange markets, the power-law exponent decreases with system size, similar to ANDROID. In other markets, there exists a non-zero correlation between system size and power-law exponent.

We systematically reveal the size dependent distribution for participant activities in three steps. First, we empirically a fit power-law distribution to the activities of participants in a month, for each month, for each StackExchange market. We follow the standard procedure to fit a power-law distribution [2]. We observe that the power-law well describes the monthly activity distributions. Second, we plot the exponents of the power-law against the number of participants for all observed months in a market, for each market in StackExchange. We observe that for most StackExchange markets, the power-law exponent decreases as the system size increases. Third, we apply linear regression to reveal the relationship between power-law exponent and system size. We observe that in general power-law exponent is negatively correlated with system size. This negative correlation is strongly visible in big knowledge markets that have at least 500 monthly participants in each month.

In Figure 3 we present empirical evidence of the size dependent distribution for answer generation in three StackExchange markets: ANDROID, APPLE, and BIOLOGY. We choose these examples to cover three possible visibilities of the size dependent distribution, as manifested by the correlation between the power-law exponent and system size—strong correlation ($|r^2| \geq 0.5$), moderate correlation ($0.3 \leq |r^2| < 0.5$), and weak correlation ($|r^2| < 0.3$).

Subject POV. The distribution of subject POV implicitly drives a market's return in terms of content dependency, as manifested by the corresponding exponent. Subject POV refers to the number of distinct perspectives on a particular content (e.g., questions) that imposes a conceptual limit to the number of dependent contents (e.g., answers). For example, an open-ended question such as 'What's your favorite book?' has many possible answers, whereas a close-ended question such as 'What's the solution for 3x+5 = 2?' has a single correct answer. In reality, most questions are neither completely open-ended nor completely closed; however, from an answerer's perspective, there's a diminishing utility in answering a question that already has an answer. This diminishing utility varies from question to question—questions asking for recommendations attract many answers, whereas questions seeking factual information attract few answers.

7.3 Uncovering the Stable Core

We uncover a stable user community in each StackExchange market, that maintains the *dynamic equilibria*—the increase or decrease in overall user community does not affect the Cobb-Douglas models. We assert that this stable user community generates a large fraction of high-threshold contents that require more effort, e.g, answers and comments, whereas the remaining users are unstable and contribute a small fraction of high-threshold contents.

We reveal the presence of the stable core by summarizing the answer contribution of users with different tenure levels (# of active months). First, for each StackExchange market, we apply equal-width binning to categorize its users into five tenure levels. Then, we plot the distribution of monthly answer contribution by the users of each category using a letter-value plot. We present the letter-value plots for ANDROID in Figure 4. We observe that monthly answer contribution is an increasing function of tenure level—the users who contribute for many months also contribute a large number of answers.
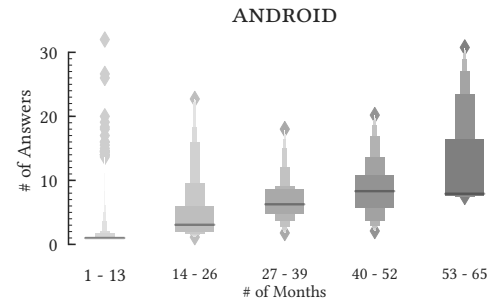
Figure 4: The distribution of monthly answer contribution of users with different # of active months, for ANDROID. The users who contribute for many months also contribute a large number of answers.

8 FAILURES AT SCALE

In this section, we discuss how and why knowledge markets may fail at scale. We first empirically examine diseconomies of scale (Section 8.1), then analyze the effects of scale on market health (Section 8.2), and finally study user exchangeability under scale changes (Section 8.3).

8.1 Diseconomies of Scale

First, we examine diseconomies of scale—the ratio of answers to questions declining with the increase in number of users. The opposite of diseconomies is economies, when the ratio increases with the increase in number of users. The concept of diseconomies is important because a decrease in the answer to question ratio implies an increase in the gap between market supply (answer) and demand (question). In fact, if the ratio falls below 1.0, the gap becomes critical—guaranteeing there will be some questions with no answers.

In Figure 5 we present the economies and diseconomies of scale in three StackExchange markets: CSTHEORY, PUZZLING and SUPE-RUSER. We choose these examples to cover three cases: strong diseconomies, strong economies, and weak economies. Among the three markets, SUPERUSER shows strong diseconomies of scale: if
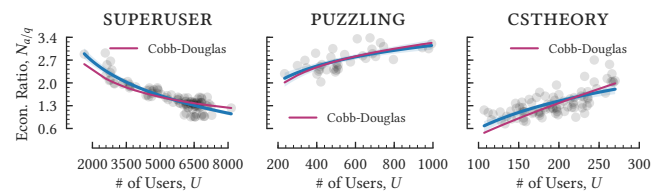
Figure 5: Diseconomies/economies of scale: the ratio of answers to questions decreasing/increasing with the increase in number of users. Most StackExchange markets exhibit diseconomies of scale. Examples: strong diseconomies—SUPERUSER; weak economies—PUZZLING; and strong economies—CSTHEORY.

the number of users increases by 1%, then the answer to question ratio declines by 0.95%. The other two markets show economies of scale, where CSTHEORY shows strong economies: if the number of users increases by 1%, then the answer to question ratio increases by 0.8%; and PUZZLING shows weak economies: if the the number of users increases by 1%, then the answer to question ratio increases by 0.2%. Note that most markets, especially the ones with more than 500 monthly active participants, exhibit diseconomies of scale similar to SUPERUSER. Only five markets exhibit strong economies of scale in StackExchange: CSTHEORY, EXPRESSIONENGINE, PUZZLING, JA_STACKOVERFLOW, and SOFTWAREENGINEERING.

The Cobb-Douglas curves well fit the empirical trends of economies and diseconomies (as shown in Figure 5). We derive these curves by dividing the answer models by the corresponding question models, and subsequently developing curves that capture economies and diseconomies ($N_{a/q}$) as a function of number of users (system size). We get similar curves via log regression. Between the two model types, the Cobb-Douglas models provide better explanation.

The Cobb-Douglas models well explain the economies and diseconomies of scale. As per the models, the primary cause of diseconomies is the difference between the diminishing returns of questions and answers for user participation. In other words, in most markets, for user input, the marginal question output is higher compared to the marginal answer output, i.e., an average user is likely to ask more questions and provide few answers. This causes the ratio of answers to questions to decline with an increase in the number of users.

8.2 Analyzing Health

Next, we examine the disadvantage of scale through two health metrics: H_1—the fraction of answered questions (questions with at least one answer); and H_2—the fraction of questions with an accepted answer (questions for which the asker marked an answer as accepted). H_1 and H_2 capture the true gap between market supply (answers) and demand (questions). An increase in the number of users may cause a decline in H_1 and H_2, as both metrics are related to the ratio of answers to questions. In fact, if the ratio falls below 1.0, it guarantees the decline of both metrics.

In Figure 6 we present the health advantage and disadvantage of scale (through H_1 and H_2) for three StackExchange markets: CSTHEORY, PUZZLING and SUPERUSER. We observe that the results are consistent with our analysis of economies and diseconomies— CSTHEORY exhibits health advantage at scale, PUZZLING remains stable, whereas SUPERUSER exhibits disadvantage at scale. These three examples cover the possible health effects of scale in knowledge markets.

8.3 Effects on Exchangeability

Finally, we empirically study the effects of scale on user exchangeability. By exchangeability, we specifically mean the gap between the top contributors and other participants in a knowledge market. Studying this gap is important because it can reveal if a market's success or failure depends on a small group of users.

To empirically study user exchangeability, we define two metrics that reflect the gap between the top contributors and other participants in a knowledge market. Note that, we only consider the
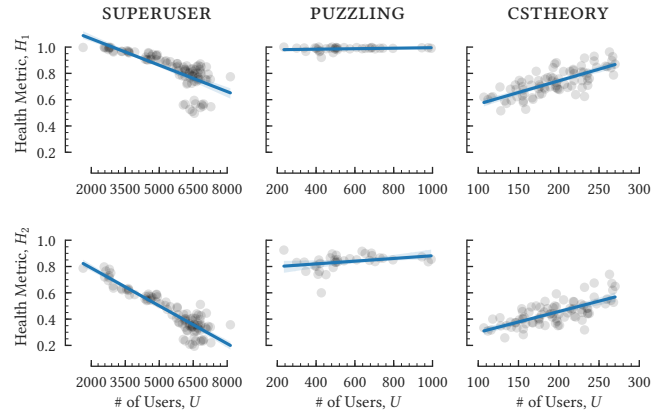
Figure 6: Health disadvantage/advantage of scale: H_1—the fraction of answered questions, and H_2—the fraction of questions with accepted answer, decreasing/increasing with the increase in number of users. Most StackExchange markets exhibit health disadvantage at scale. Examples: disadvantage—SUPERUSER; neutral—PUZZLING; and advantage—CSTHEORY.

active participants who contributed at least one content. The first metric E_1 is defined as the ratio of contribution between the top 5% and the bottom 5% of users. For computing E_1, we measure the contribution of a user v as the ratio $N_{a/q}^v$ of the number of answers N_a^v provided by the user to the number of questions N_q^v asked by the user. Notice that E_1 is a ratio based metric and we define user contribution to be consistent with this metric. The second metric E_2 is defined as the sum of two distances: (i) the distance between
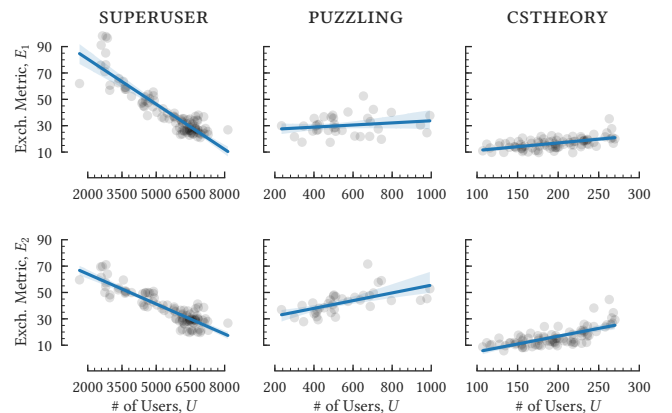
Figure 7: User exchangeability under scale: the gap (E_1 and E_2) between the top contributors and other participants in a knowledge market decreasing or increasing with the increase in number of users. Most markets exhibit a large gap between the top contributors and other participants. Examples: high dissimilarity—SUPERUSER; moderate dissimilarity—PUZZLING; and low dissimilarity—CSTHEORY.

the contribution of the top 5% of users and the median 5% of users, and (ii) the distance between the contribution of the median 5% of users and the bottom 5% of users. For computing E_2, we measure the contribution of a user v as a tuple $< N_a^v, N_q^v >$, consisting of the number of answers N_a^v provided by the user and the number of questions N_q^v asked by the user. Notice that E_2 is an interval based metric and we define user contribution to be consistent with this metric. While both metrics have certain limitations, e.g., they are sensitive to outliers, these metrics allow us to comprehend user exchegeability to some extent.

In Figure 7 we present the exchangeability of users under scale changes (through E_1 and E_2) for three StackExchange markets: CSTHEORY, PUZZLING and SUPERUSER. Among the three markets, SUPERUSER exhibits the highest gap between the top contributors and the other participants. However, as the number of participants increases, this gap decreases, i.e., the users become more exchangeable. In contrast, CSTHEORY exhibits the lowest gap between the top contributors and the other participants. However, as the number of participants increase, this gap increases, i.e., the users become less exchangeable.

9 IMPLICATIONS

Our work promotes two new research directions—size-dependent mechanism design and content dependency in social media—while advancing several others—metrics of market health, power law of participation, and microfoundations of knowledge markets.

Size-Dependent Mechanism Design. We reveal that the health of a knowledge market depends on the market's size. A natural implication of this dependency is that site operators should adjust mechanisms based on the number of participants. For example, a site operator can decide between retaining existing users (via incentives) and attracting new users (via advertising) based on the number of participants and their activity distribution.

Content Dependency in Social Media. We observe that many social media platforms support several possible user actions with "complex dependencies". For example, in Facebook, a post is the root content (primary), comments on the post nest below the post (secondary), and replies to these comments nest beneath the original comments (tertiary). Further, a user can react to any of these content types with several possible reactions. Overall user activity in Facebook is distributed across these possible actions with complex dependencies, which drives the platform's health.

Metrics of Market Health. We demonstrate the presence of diseconomies of scale with several metrics that partially capture the health of a knowledge market. While we concentrate on content-generation based *production metrics*, our concepts can be extended for page-view based *consumption metrics* as well. Also, there is room for developing new health metrics that capture a more detailed picture of a knowledge market's health including *market efficiency*—the degree to which market price (amount of responses and reactions) is an unbiased estimate of the true value of the investment (user effort in content generation) [11].

Power Law of Participation. In StackExchange markets, a small fraction of the user community participate in high-engagement activities (e.g., linking similar questions), whereas the larger fraction participate in low-threshold activities (e.g., voting). This asymmetry leads to a *Power Law of Participation* [29]. We assert that both low-threshold and high-engagement activities are required for a knowledge market's survival, and should proportionately increase with the increase in number of participants. However, in reality, for most knowledge markets, the size of the user community contributing high-engagement activities does not scale with the system size. This creates a "gap" between market supply and demand, and consequently affects market health.

Microfoundations of Knowledge Markets. The size-dependent distribution of user contribution implies that users who join a community later in its lifecycle exhibit different behavior than those who were present from the beginning. This very well may imply that the distribution of individual user behaviors (not just their overall production) is "also" a function of the system size. We should expect to see a stable user behavior distribution over time for markets that appear to be more scale-insensitive; preliminary results suggest that this may indeed be the case [17].

10 LIMITATIONS

We discuss several limitations of our work. First, the economic production models do not account for user growth. While there exist several user growth models for two-sided markets [24], membership based websites [32], and online social networks [6, 21, 45], it would be useful to introduce economic user growth models that complement our proposed content generation models. Specifically, there is a need to develop resource-based user growth models that account for market health. A potential direction in this research is to extend the Malthusian growth model [28]. Second, the proposed production models inherit the fundamental assumptions of macroeconomics: an aggregate is homogeneous (without looking into its internal composition), and aggregates are functionally related etc. [33]. It would be useful to empirically study these assumptions for real-world knowledge markets.

11 CONCLUSION

In this paper, we examined the CQA websites on StackExchange platform through an economic lens by modeling them as knowledge markets. In particular, we designed a set of production models to capture the content generation dynamics in these markets. The resulting best-fit model, Cobb-Douglas, predicts the production of content in StackExchange markets with high accuracy. We showed that the model provides intuitive explanations for content generation. Specifically, it reveals that factors of content generation such as user participation and content dependency have *constant elasticity*; in many markets, factors exhibit *diminishing returns*; markets vary according to their *returns to scale*; and finally many markets exhibit *diseconomies of scale*. We further investigated these prognoses by showing the presence of diseconomies of scale in terms of content production, and several measures of market health. The implications of our work are two-fold: site operators need to design incentives as a function of number of participants; there is a need to develop Economic lenses that can shed insights into the complex dependencies amongst different content types and participant actions in general social networks.

REFERENCES

[1] Mohammed Abufouda. 2017. Community Aliveness: Discovering Interaction Decay Patterns in Online Social Communities. *CoRR* abs/1707.04477 (2017). arXiv:1707.04477 http://arxiv.org/abs/1707.04477

[2] Lada A Adamic. 2000. Zipf, power-laws, and pareto-a ranking tutorial. *Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html* (2000).

[3] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proceedings of the 17th International Conference on World Wide Web (WWW '08)*. ACM, New York, NY, USA, 665–674. https://doi.org/10.1145/1367497.1367587

[4] Pal Aditya, Chang Shuo, and Konstan Joseph, A. 2012. Evolution of Experts in Question Answering Communities. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media.*

[5] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. 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 (KDD '12)*. ACM, New York, NY, USA, 850–858. https://doi.org/10.1145/2339530.2339665

[6] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. 2006. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. ACM, New York, NY, USA, 44–54. https://doi.org/10.1145/1150402.1150412

[7] Ricardo Baeza-Yates and Diego Saez-Trumper. 2015. Wisdom of the Crowd or Wisdom of a Few?: An Analysis of Users' Content Generation. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15)*. ACM, 69–74. https://doi.org/10.1145/2700171.2791056

[8] Mary Ann Branch, Thomas F Coleman, and Yuying Li. 1999. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM Journal on Scientific Computing* 21, 1 (1999), 1–23.

[9] Brian S Butler. 2001. Membership size, communication activity, and sustainability: A resource-based model of online social structures. *Information systems research* 12, 4 (2001), 346–362.

[10] Yan Chen, TECK-HUA HO, and YONG-MI KIM. 2010. Knowledge market design: A field experiment at Google Answers. *Journal of Public Economic Theory* 12, 4 (2010), 641–664.

[11] Aswath Damodaran. 2002. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset.* Wiley.

[12] Katherine Ellis, Moises Goldszmidt, Gert Lanckriet, Nina Mishra, and Omer Reingold. 2016. Equality and Social Mobility in Twitter Discussion Groups. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16)*. ACM, New York, NY, USA, 523–532. https://doi.org/10.1145/2835776.2835814

[13] Desta Fekedulegn, Siútáin Mártín Pádraig Mac, and Jim J. Colbert. 1999. Parameter Estimation of Nonlinear Models in Forestry. *Finnish Society of Forest Science and the Finnish Forest Research Institute* (1999).

[14] Emilio Ferrara, Nazanin Alipourfard, Keith Burghardt, Chiranth Gopal, and Kristina Lerman. 2017. Dynamics of Content Quality in Collaborative Knowledge Production. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.* 520–523.

[15] Adabriand Furtado, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. 2013. Contributor Profiles, Their Dynamics, and Their Importance in Five Q&A Sites. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, 1237–1252. https://doi.org/10.1145/2441776.2441916

[16] David Garcia, Pavlin Mavrodiev, and Frank Schweitzer. 2013. Social Resilience in Online Communities: The Autopsy of Friendster. In *Proceedings of the First ACM Conference on Online Social Networks (COSN '13)*. ACM, New York, NY, USA, 39–50. https://doi.org/10.1145/2512938.2512946

[17] Chase Geigle, Himel Dev, Hari Sundaram, and Chengxiang Zhai. 2018. *Discovering and Analyzing Action-Based Roles in Community Question Answering Networks.* Technical Report.

[18] Glossary of economics. 2017. Glossary of economics — Wikipedia, The Free Encyclopedia. (2017). https://en.wikipedia.org/wiki/Glossary_of_economics [Online; accessed 30-October-2017].

[19] Benjamin V. Hanrahan, Gregorio Convertino, and Les Nelson. 2012. Modeling Problem Difficulty and Expertise in Stackoverflow. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion (CSCW '12)*. ACM, New York, NY, USA, 91–94. https://doi.org/10.1145/2141512.2141550

[20] Heike Hofmann, Hadley Wickham, and Karen Kafadar. 2017. Letter-Value Plots: Boxplots for Large Data. *Journal of Computational and Graphical Statistics* 26, 3 (2017), 469–477. https://doi.org/10.1080/10618600.2017.1305277

[21] Sanjay Ram Kairam, Dan J. Wang, and Jure Leskovec. 2012. The Life and Death of Online Groups: Predicting Group Growth and Longevity. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. ACM, New York, NY, USA, 673–682. https://doi.org/10.1145/2124295.2124374

[22] Komal Kapoor, Mingxuan Sun, Jaideep Srivastava, and Tao Ye. 2014. A Hazard Based Approach to User Return Time Prediction. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, New York, NY, USA, 1719–1728. https://doi.org/10.1145/2623330.2623348

[23] Robert E. Kraut and Andrew T. Fiore. 2014. The Role of Founders in Building Online Groups. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 722–732. https://doi.org/10.1145/2531602.2531648

[24] Ravi Kumar, Yury Lifshits, and Andrew Tomkins. 2010. Evolution of Two-sided Markets. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*. ACM, New York, NY, USA, 311–320. https://doi.org/10.1145/1718487.1718526

[25] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2006. Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. ACM, New York, NY, USA, 611–617. https://doi.org/10.1145/1150402.1150476

[26] Varun Kumar and Niranjan Pedanekar. 2016. Mining Shapes of Expertise in Online Social Q&A Communities. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '16 Companion)*. ACM, New York, NY, USA, 317–320. https://doi.org/10.1145/2818052.2869096

[27] Zhiyuan Lin, Niloufar Salehi, Bowen Yao, Yiqi Chen, and Michael Bernstein. 2017. Better When It Was Smaller? Community Content and Behavior After Massive Growth. In *Proceedings of the 11th International AAAI Conference on Web and Social Media.*

[28] Thomas Robert Malthus. 1809. *An essay on the principle of population, as it affects the future improvement of society.* Vol. 2.

[29] Ross Mayfield. 2006. Power Law of Participation. (2006). http://ross.typepad.com/blog/2006/04/power_law_of_pa.html [Online; accessed 30-October-2017].

[30] Akshay Patil, Juan Liu, and Jie Gao. 2013. Predicting Group Stability in Online Social Networks. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, 1021–1030. https://doi.org/10.1145/2488388.2488477

[31] Liu Qiaoling and Agichtein Eugene. 2011. Modeling Answerer Behavior in Collaborative Question Answering Systems. In *Proceedings of the 33rd European conference on Advances in information retrieval.*

[32] Bruno Ribeiro. 2014. Modeling and Predicting the Growth and Death of Membership-based Websites. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, New York, NY, USA, 653–664. https://doi.org/10.1145/2566486.2567984

[33] Christopher A Sims. 1980. Macroeconomics and reality. *Econometrica: Journal of the Econometric Society* (1980), 1–48.

[34] Ivan Srba and Maria Bielikova. 2016. A Comprehensive Survey and Classification of Approaches for Community Question Answering. *ACM Trans. Web* 10, 3, Article 18 (Aug. 2016), 63 pages. https://doi.org/10.1145/2934687

[35] Ivan Srba and Maria Bielikova. 2016. Why is Stack Overflow Failing? Preserving Sustainability in Community Question Answering. *IEEE Software* 33, 4 (July 2016), 80–89. https://doi.org/10.1109/MS.2016.34

[36] Jim Stanford. 2008. Economics for Everyone: On-Line Glossary of Terms & Concepts. (2008).

[37] David Tilman. 1980. Resources: A Graphical-Mechanistic Approach to Competition and Predation. *The American Naturalist* 116, 3 (1980), 362–393. https://doi.org/10.1086/283633 arXiv:https://doi.org/10.1086/283633

[38] Utkarsh Upadhyay, Isabel Valera, and Manuel Gomez-Rodriguez. 2017. Uncovering the Dynamics of Crowdlearning and the Value of Knowledge. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. ACM, 61–70. https://doi.org/10.1145/3018661.3018685

[39] Simon Walk, Denis Helic, Florian Geigl, and Markus Strohmaier. 2016. Activity Dynamics in Collaboration Networks. *ACM Trans. Web* 10, 2, Article 11 (May 2016), 32 pages. https://doi.org/10.1145/2873060

[40] Gang Wang, Konark Gill, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. 2013. Wisdom in the Social Crowd: An Analysis of Quora. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, 1341–1352. https://doi.org/10.1145/2488388.2488506

[41] Lingfei Wu, Jacopo A. Baggio, and Marco A. Janssen. 2016. The Role of Diverse Strategies in Sustainable Knowledge Production. *PLoS ONE* (2016).

[42] Lingfei Wu and Jiang Zhang. 2011. Accelerating growth and size-dependent distribution of human online activities. *Phys. Rev. E* 84 (Aug 2011), 026113. Issue 2. https://doi.org/10.1103/PhysRevE.84.026113

[43] Jie Yang, Alessandro Bozzon, and Geert-Jan Houben. 2015. "Harnessing Engagement for Knowledge Creation Acceleration in Collaborative Q&A Systems". In *Proceedings of the 23rd International Conference on user Modeling, Adaptation, and Personalization (UMAP)*, Francesco Ricci, Kalina Bontcheva, Owen Conlan, and Séamus Lawless (Eds.). 315–327.

[44] Tausczik Yla, Wang Ping, and Joohee Choi. 2017. Which Size Matters? Effects of Crowd Size on Solution Quality in Big Data Q&A Communities. In *Proceedings of the 11th International AAAI Conference on Web and Social Media.*

[45] Chengxi Zang, Peng Cui, and Christos Faloutsos. 2016. Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining (KDD '16). ACM, New York, NY, USA, 2015–2024. https://doi.org/10.1145/2939672.2939825

[46] Jun Zhang, Mark S. Ackerman, and Lada Adamic. 2007. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 221–230. https://doi.org/10.1145/1242572.1242603

[47] Haiyi Zhu, Jilin Chen, Tara Matthews, Aditya Pal, Hernan Badenes, and Robert E. Kraut. 2014. Selecting an Effective Niche: An Ecological View of the Success

of Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 301–310. https://doi.org/10.1145/2556288.2557348

[48] Haiyi Zhu, Robert E. Kraut, and Aniket Kittur. 2014. The Impact of Membership Overlap on the Survival of Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 281–290. https://doi.org/10.1145/2556288.2557213