# Jointly Modeling Population and Content Growth in Stack Exchange Websites

Himel Dev, Chase Geigle, Hari Sundaram
Department of Computer Science
University of Illinois at Urbana-Chamapaign (UIUC)
hdev3,geigle1,hs1@illinois.edu

**ACM Reference format:**
Himel Dev, Chase Geigle, Hari Sundaram. 2017. Jointly Modeling Population and Content Growth in Stack Exchange Websites. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 2 pages.
https://doi.org/10.475/123_4

## 1 MOTIVATION

Interacting population based models find use in a variety of domains to model population growth. In recent times, there has been an increasing interest in applying such models to explain the growth or the decline in online social networks [1, 2]. While these existing models have been successful in capturing population growth,

I would rewrite this sentence in the following manner:
While these interacting population models have been successful in capturing social network growth, they have limitations. These models are not expressive enough to incorporate latent factors; lack model interpretability; account for social network actor presence or absence, but not the content that they produce while they are on the social network.

they have some fundamental limitations such as incorporating largely unobservable factor(s), lack of interpretation in explaining growth

Is lack of interpretability also true of Ribeiro [2]? I thought the $\frac{\beta}{\alpha}$ value helped explain the growth of the network.

, not incorporating observable content factors.

Kumar et al. [1] introduced a two-sided (two types of users: blue and green) market model based on the hypothesis that a blue user decides to join a platform based on the platforms' current proportion of green users and vice versa. However, in most social networks, only a small fraction of user community is visible to a potential user, and the proportion of a particular type of user in the network remains largely unobservable.

why can we not use the theory of two sided markets to explain content production and participation?

Ribeiro et al. [2] proposed a daily active user (DAU) prediction model based on a set of reaction-diffusion-decay equations describing the interactions between active members, inactive members, and not-yet-members. The proposed model captures the number of daily active users using a set of growth parameters. However, these growth parameters provide little insight in understanding the actual reasons behind the growth/decline in DAU.

we have to be careful in how we review Ribeiro [2]; the paper states that the $\beta$ and $\alpha$ parameters are interpretable; we need a sharp contrast here since we will also use interaction population models

We observe that while population interaction can model population growth in certain social networks, it provides little insight in explaining growth.

Is the problem with the model (i.e. not expressive enough; not interpretable), or is that the model when used for a social network does not incorporate content?

We hypothesize that population content interaction is a more effective means to model and explain growth in both population and content. Growth models based on population content interaction can reveal the interdependency between the two, i.e., how growth/decline in one affects the other. Such insights are valuable for content based networks, in particular Q&A networks. In this paper, we focus on jointly modeling population and growth in Stack Exchange websites, revealing the interdependency between the two.

## 2 POPULATION CONTENT INTERACTION

We present a set of population content inductions/reactions

You need to have a section that introduces the idea of reactions

to capture the interaction between population and content in Stack Exchange websites.

changed the style of the table to make it look "clean"; to denote comments conditioned on questions, should we use $\Delta N(c|q;t)$ instead? Also $U(c;t) = U(c|q;t) + U(c|a;t)$ is straightforward to interpret.

**I. Answer Reaction:** In a Stack Exchange website, there are two key factors in generating answers: questions,

how do you define an active question?

and users who answer questions (aka answerers). Based on these two factors, we assert the following reaction to generate answers.

What do the coefficients on the left side of the equation mean? In the equation below, the number of active questions are getting depleted; but then, the meaning of the right side is the number of accepted answers; else an active question cannot become inactive; the other way to interpret this equation is to think the left side has number of unanswered questions (i.e. questions without any answers), which are getting depleted at rate $\alpha_{1,1}$.

$$\alpha_{1,1}\Delta N_q + \alpha_{2,1}U_a \implies \beta_{1,1}\Delta N_a$$

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
*Conference'17, July 2017, Washington, DC, USA*
© 2016 Copyright held by the owner/author(s).
ACM ISBN 123-4567-24-567/08/06…$15.00
https://doi.org/10.475/123_4

| Symbol | Interpretation |
|---|---|
| $U_q(t)$ | # 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 $t$ |
| $\Delta N_q(t)$ | # of active questions at time $t$ |
| $\Delta N_a(t)$ | # of answers to active questions at time $t$ |
| $\Delta N_c^q(t)$ | # of comments to active questions at time $t$ |
| $\Delta N_c^a(t)$ | # of comments to active answers at time $t$ |
| $\Delta N_{+v}^q(t)$ | # of upvotes to active questions at time $t$ |
| $\Delta N_{+v}^a(t)$ | # of upvotes to active answers at time $t$ |
| $\alpha_{i,j}$ | Coefficient of $i$th input in $j$th induction/reaction |
| $\beta_{i,j}$ | Coefficient of $i$th output in $j$th induction/reaction |

**Table 1: Notations in inductions/reactions**

**II. Question Reaction:** In a Stack Exchange website, there is a single key factor in generating questions: users who ask questions (aka askers). Based on this single factor, we assert the following reaction to generate questions.

$$\alpha_{1,2} U_q \Longrightarrow \beta_{1,2} \Delta N_q$$

**III. Answerer Induction:** In a Stack Exchange website, there are two key factors in inducing the number of answerers at time $t$: number of answerers at time $(t-1)$,

Do we need to define reaction equations using variables $t$ and $t-1$? It is implicit in the reaction definition. I don't remember seeing this in the Ribeiro [2] paper.

and the utility received by these answerers at time $(t-1)$. Now, in a simplistic model the utility received by the answerers at time $(t-1)$ can be captured using the number of active questions at time $(t-1)$. Based on these factors, we assert the following induction to induce the number of answerers at time $t$.

$$\alpha_{1,3} U_a(t-1) + \alpha_{2,3} \Delta N_q(t-1) \Longrightarrow \beta_{1,3} U_a(t)$$

An extended interaction model where answerers' utility is defined in terms of questions, comments to answers, and upvotes to answers is as follows.

$$\left.\begin{aligned} &\alpha_{1,3} U_a(t-1) + \alpha_{2,3} \Delta N_q(t-1) + \\ &\alpha_{3,3} \Delta N_c^a(t-1) + \alpha_{4,3} \Delta N_{+v}^a(t-1) \end{aligned}\right\} \Longrightarrow \beta_{1,3} U_a(t)$$

**IV. Asker Induction:** In a Stack Exchange website, there are two key factors in inducing the number of askers at time $t$: number of askers at time $(t-1)$, and the utility received by these askers at time $(t-1)$. Now, in a simplistic model the utility received by the askers at time $(t-1)$ can be captured using the number of active answers at time $(t-1)$. Based on these factors, we assert the following reaction to induce the number of askers at time $t$.

$$\alpha_{1,4} U_q(t-1) + \alpha_{2,4} \Delta N_a(t-1) \Longrightarrow \beta_{1,4} U_q(t)$$

An extended interaction model where askers' utility is defined in terms of answers, comments to questions, and upvotes to questions is as follows.

$$\left.\begin{aligned} &\alpha_{1,4} U_q(t-1) + \alpha_{2,4} \Delta N_a(t-1) + \\ &\alpha_{3,4} \Delta N_c^q(t-1) + \alpha_{4,4} \Delta N_{+v}^q(t-1) \end{aligned}\right\} \Longrightarrow \beta_{1,4} U_q(t)$$

## 3 GROWTH MODEL

A production function captures the relationship between the output of a production process, and the inputs or factors of production.

Why do we need two different kinds of relationships? We seem to have a model based on reactions and another based on economics; how do we know that these are consistent in some sense; that is, these relationships don't work cross purposes.

We use the Cobb-Douglas production function to capture the relationship between the output, and the inputs or factors in our inductions/reactions. The Cobb-Douglas function is of following form.

$$Y = A \prod_{i=1}^{n} X_i^{\lambda_i}$$

Here, $Y$ represents total production at time $t$, $X_i$ represents total $i$th input at time $t$, $\lambda_i$ represents output elasticity of the $i$th input, and $A$ represents total factor productivity. Based on our population content inductions/reactions and the aforementioned production function, we derive the following set of relationships

Can you justify each equation?

between population and content.

$$N_a = A_1 N_q^{\lambda_{1,1}} U_a^{\lambda_{2,1}} \tag{1}$$

$$N_q = A_2 U_q^{\lambda_{1,2}} \tag{2}$$

$$U_a(t) = A_3 [U_a(t-1)]^{\lambda_{1,3}} [N_q(t-1)]^{\lambda_{2,3}} \tag{3}$$

$$U_q(t) = A_4 [U_q(t-1)]^{\lambda_{1,4}} [N_q(t-1)]^{\lambda_{2,4}} \tag{4}$$

In addition to these relationships, we incorporate resource constraints on our growth model. More specifically, in a Stack Exchange website, every user $u$ has a fixed resource capacity $r_u$ at a given time, which he/she can use to ask questions, answer questions, and make comments. We assume voting activity consumes negligible resources. Now, the resource capacity distribution $P_r$ can be defined as the probability that a user–chosen uniformly at random from the set of all users–has resource capacity $c$. The mean resource capacity $z$ can be calculated as $z = \sum_r r P_r$. To incorporate resource constraint in our growth model in a simplistic manner, we hold the assumption that every user has the mean resource capacity $z$. This assumption is based on the mean field approximation, which allows us to focus on the overall resource capacity of the entire population.

$$\Delta N_q(t) + \Delta N_a(t) + \Delta N_c(t) = zU(t) \tag{5}$$

## REFERENCES


[1] 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

[2] 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