Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

A MODEL OF ONLINE COLLABORATION FOR KNOWLEDGE PRODUCTION

Miles K. Manning

School of Human Evolution
and Social Change
Arizona State University
900 Cady Mall
Tempe, AZ 85281, USA

Marco A. Janssen

School of Sustainability
Arizona State University
800 Cady Mall
Tempe, AZ 85281, USA

Lingfei Wu

Computation Institute
University of Chicago
5735 S. Ellis Ave
Chicago, IL 60607, USA

ABSTRACT

Large scale collaboration is a fundamental characteristic of human society, and has recently manifested in the development and proliferation of online communities. These virtual social spaces provide an opportunity to explore large scale collaborations as natural experiments in which determinants of success can be tested. In order to do this, we first review previous work on meddling online communities to build an understanding of how these communities function. Having thus identified the operating mechanisms inherent in online communities, we propose a population ecology model of online communities that seeks to explain a number of statistical patterns from a selection of such communities.

1 INTRODUCTION

Throughout history, large scale collaboration has allowed for the foundation of cities, trade networks, scientific advancement, medical breakthroughs and the the advancement of human rights. The most visible of these collaborations are those that succeeded, as failed collaborations do not leave much in the way of artifacts or historical records. Thus, exploration of what determines success in collaboration is necessarily limited to evaluating success, rather than learning from failure.

However, the nature of the internet has lowered the cost of collaboration by allowing large groups to meet and share information with unprecedented ease and speed. This ease of exchange has resulted in notable collaborative outcomes such as Wikipedia, Stack Exchange and Linux. Thus, online communities have provided an unprecedented source of data for learning what enables collaborations to succeed through an examination of failed collaborations: i.e., they start, proliferate, fail and disappear rapidly and often in a public sphere.

In order to use online communities as a data source for understanding large scale collaboration, the nature of these communities must first be understood. Online communities are, of course, social networks of exchange that are founded in and operate in online spaces. They can be seen as an environment shaped by users within these spaces, and can take on many varied forms. The tasks users complete further the

goals of the community and are the resource of the environment. These tasks are provided by communities and consumed by users. There is not enough attention available to make all online communities vibrant and productive; this means that in order to better understand these spaces, it is prudent to model the user population rather than individual project. Communities harvest attention from the user population and use it to create knowledge products. Successful collaborations are those that are able to attract the attention of a population sufficient to allow that community to remain productive.

Before constructing a model, a review of prior work is used to identify what patterns exist across online communities, as well as what mechanisms might drive the behavior of online collaborators. Subsequently, an agent-based model of online communities is described and the simulation results are compared to a variety of patterns found in empirical studies of online communities.

2 PATTERNS IN ONLINE COMMUNITIES

Before constructing a model, a review of previous publications is performed to determine what key patterns exist in online communities (Table 1). Although online communities have data available, this data has limitations that must be considered. Browsing is an important example of such limitations, for in browsing activity data comes from clicks, edits, and contributions. Depending on the specific source of data, browsing might be indistinguishable from contributions, or might be undetectable. The web crawl is another confounding factor, as it provides information about the structure of the web and can inform as to the relative size of websites while revealing nothing about user traffic. Though the logged data of online communities is accessible and viable, it is important to be cognizant of such limitations.

Web crawlers can provide information on the structure of the web, and provide the first look data for understanding the environment in which online communities live. If web pages are considered as nodes and hyperlinks as edges, both the in-degree and out-degree of blogs obey a power-law distribution (Fu, Liu, and Wang 2008). Further, both the web (Barabási and Albert 1999) and the internet (Albert and Barabási 2002) have a degree distribution that is power-law. Lastly, the size of both strongly and weakly, connected components, are also power-law distributed (Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, and Wiener 2000).

A key source of information for modeling collaborative communities is data collected from some exemplar communities. For example, Wikis are a popular platform for online collaboration, as well as an excellent source of data. The number of editors a wiki has appears to be distributed according to a Zipf law, as does the number of edits until falling below a critical mass (Kittur and Kraut 2010). Furthermore, the distribution of Wikipedia editors per article appears to follow a power-law distribution for most of the range (Yasseri and Kertész 2013).

The next pattern of interest is the behavior of users. First, the number of clicks made while browsing the internet is distributed according to a power law (Huberman, Pirolli, Pitkow, and Lukose 1998). More relevantly, user contribution seems to be driven by attention (Wu, Wilkinson, and Huberman 2009). While this does not always result in a power-law, the distribution seems to consistently have a high probability of few contributions, and a power law on a portion of the domain (Wilkinson 2008, Wu, Wilkinson, and Huberman 2009, Ozmen, Smith, Yilmaz, and Smith 2012). The slope of this power law is dependent on the community (Wilkinson 2008), as is the domain (Wu, Wilkinson, and Huberman 2009).

Patterns in the literature focus on some basic power laws that could be explained by various mechanisms. To derive more specific empirical patterns to challenge the model, it will be compared to an analysis of Stack Exchange, a network of question-and-answer communities covering diverse topics in a variety of different fields. The Stack Exchange network is meant to provide users with expert answers, a goal accomplished by establishing sub-communities that are specialized in different topics and sub-topics. The Stack Exchange community was chosen for in-depth analysis for several important reasons. Stack Exchange contains enough communities that the model can be compared to distributions of their characteristics rather than specific examples, while also providing a well-defined subset of the web to model. The focus of Stack Exchange

Table 1: Characteristics of various online communities and their sources.

Reference	**Subject**	**Users per Community**	**Tasks per User**	**Findings**
(Wilkinson 2008)	Peer-Production	$2-4$	$10-30$	scale free distribution of contributions per user slope of distribution task dependent scale free distribution of contributions per community
(Fu, Liu, and Wang 2008)	Blogging	1	NA	power law in-degree power law out-degree
(Wu, Wilkinson, and Huberman 2009)	Digg/YouTube	NA	NA	attention motivates production power law user contribution in Digg long tail user contribution in YouTube
(Huberman, Pirolli, Pitkow, and Lukose 1998)	Browsing	NA	NA	power law browsing behavior
(Huberman and Adamic 1999)	Website	NA	NA	power law size of website
(Albert and Barabási 2002)	Internet	NA	NA	power law degree distribution
(Barabási and Albert 1999)	WWW	NA	NA	power law degree
(Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, and Wiener 2000)	WWW	NA	NA	power law weakly connected components power law strongly connected components
(Radtke 2011)	FLOSS	2	$20-30$	power law project developers
(Ozmen, Smith, Yilmaz, and Smith 2012)	Participatory Science	2	1	power law user contribution
(Kittur and Kraut 2010)	Wiki	10	0.2	Zipf law wiki users Zipf law wiki edits
(Yasseri and Kertész 2013)	Wikipedia	$0.25-1.25$	100	power law article editors log normal inter edit time power law session edits

on expert answers is amendable to our desire to specifically model communities of skilled users. The easy availability of Stack Exchange data was also an important factor in choosing it for analysis.

3 PREVIOUS MODELING WORK

A survey of earlier papers on modeling online communities reveals mechanisms of user action thought to be present. The search is restricted to papers that take a mechanistic approach to modeling online communities, and models of online social networks are excluded from the survey. Communities are considered productive communities if they create or distribute knowledge, or practice peer production. Examples include open source software, Wikipedia, and question and answer forums such as Stack Exchange. With these restrictions, a few different mechanisms are found that, in various forms, are used to model online communities. These three mechanisms, known as preferential attachment, foraging, and infection, apply to different aspects of online communities.

Power law distributions are a hallmark of online communities (Wilkinson 2008, Fu, Liu, and Wang 2008, Barabási and Albert 1999, Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, Tomkins, and Wiener 2000, Huberman, Romero, and Wu 2009). Preferential attachment is a mechanism for the formation of power law distributions, which has been verified to occur in online communities (Pastor-Satorras, Vázquez, and Vespignani 2001). The idea that "the rich get richer" is captured in the positive feedback of preferential attachment. If one were to apply this idea to online communities, one could say that a new user is more likely to join the larger communities than the smaller. This concept could also be applied to tasks or connections between communities. That is, communities with a large number of tasks will draw more new tasks, or highly connected communities are more likely to form new connections.

One way of explaining the time people spend contributing to online communities is to assume they derive some benefit from this activity. If contributing satisfies some need for users, they can be considered to forage in communities for tasks. One successful foraging heuristic is win-stay, lose-shift; this is when a forager categorizes their outcomes as either sufficient or not, and moves from their foraging site only when their outcomes are insufficient (Nowak and Highfield 2011, Ozmen, Smith, Yilmaz, and Smith 2012). In the language of online communities, one would consider a user remaining in a community only as long as the user is able to contribute to that community.

Another modeling framework with some applicability to online communities is epidemiology. It has been observed that popularity drives production in online communities. Considering contributors as infectious is one way of explaining or conceptualizing this phenomenon. Users could be seen to follow an SEIR (Susceptible, Exposed, Infectious, Recovered) progression, where unattached susceptible users (S) are exposed to a community or task by contributors (I), after which they are considered exposed (E). After some time, exposed users progress in turn to active contributors (I), eventually contributors (I) leave the community and begin a refractory period (R) during which they will not re-join the community (Ozmen, Smith, Yilmaz, and Smith 2012). One of the common and basic properties of infectious disease models is that the number of new infections is proportional to the number of infectious people. This corresponds to the number of active contributors in a community, thus determining its attractiveness. The above presented parallel concept highlights the importance of popularity in online communities and their self-reinforcing behavior.

From the review a few key mechanisms used in the modeling of online communities can be identified: users must follow other users; positive feedback of popularity is included in both preferential attachment and epidemiological models; contributions must be a commodity sought after by users and communities. This can be explained by the observation that attention drives contribution (Huberman, Romero, and Wu 2009, Wu, Wilkinson, and Huberman 2009). Since users follow tasks, and popularity predicts user movement, tasks must occur with higher frequency in larger communities. The next section describes a model of online communities.

Table 2: This table summarizes the variables and parameters of the model.

Variable	Description	Value
$NumComm$	Number of communities	500
$NumUser$	Number of users	500
$NumTask$	Number of tasks	500
$MinP$	Minimum task number	4
$MaxP$	Maximum task number	100
i	Community index	$i = 1,2,...,NumComm$
j	User index	$j = 1,2,...,NumUser$
k	Task index	$k = 1,2,...,NumTask$
s	User skill	$2 \leq s \leq \sqrt{MaxP}$
P_i	User population of community i	$0 \leq P_i \leq NumUser$
T_i	Topic of community i	$MinP \leq T_i \leq MaxP$
A_j	Attention level of user j	$A_j = 0,1$
$S(j)$	Skill of user j	$2 \leq S(j) \leq \sqrt{MaxP}$
UC_j	Community that user j belongs to	$UC_j \in i$
TC_k	Community that task k belongs to	$UT_k \in i$
TN_k	Task number of task k	$MinP \leq TN_k \leq MaxP$
$CH_{k,s}$	Record of skill s being applied to task k	$CH_{k,s} = 0,1$
rr	Community replacement rate	0.05
λ	Topic adjustment rate	0.1

4 MODEL DESCRIPTION

4.1 Introduction

In order to study the population dynamics of online communities, a simulation model is developed. In this model, users can move between communities while completing tasks that they encounter. The model includes: task generation, task allocation, user contribution, task completion, and user movement.

Three types of agents are considered in this model: tasks, users, and communities. Tasks are described by a task number, a list of past contributions, and a community to which that task belongs. Users are described by a skill number, their attention level, and the community they are participating in. Communities have a topic as well as user and task populations. The model is run in discrete time and the state variables are modified at each time step. Table 2 provides a complete list of variables and parameters.

The model used is a foraging model, tailored to suit online communities. Users are the foragers and they forage for tasks. However, users are heterogeneous, and a task that one user can complete might be outside the skill set of another user. Thus, tasks can be thought of as the resource users are foraging for, and users contributing to a task is what consumes that resource. Communities serve to divide the simulation environment into locations for foragers to move between.

4.2 Tasks

The task considered in this model is the identification of prime numbers. This task was chosen because it allows for heterogeneity in task "topic, as well as heterogeneity in agent skill. A user's skill is thought of as an integer. This integer is the number the user can divide by. A task is complete when a user successfully divides the task (nonprime) or when every applicable skill has been applied to the task (prime). Tasks are the resource that users forage for. During each time step of the model tasks must be generated, allocated to communities, receive user contributions, and checked for completion.

In this model task generation occurs separately from task allocation. It is assumed that tasks are generated independently of the communities modeled. This could be justified by the observation that in online communities, those that ask questions, and those that answer them, are largely disjointed. Newly generated tasks then occur according to the parameters of the model rather than the dynamic variables. When a new task is generated it is chosen randomly to be any of the possible tasks considered (integers between $MinP$ and $MaxP$).

$$P\left(TN_{\{k|TC_k(t)=0\}}(t+1)=x\right)=\left\{\begin{array}{l|l}\frac{1}{1+MaxP-MinP} & x\in[MinP,MaxP]\\ 0 & x\notin[MinP,MaxP]\end{array}\right\}$$

Created tasks must be assigned to a community. Tasks are assigned based on two key factors: the size of the community (P_i) and the topic of the community (T_j). Community size is included to mimic the effect that popular communities receive more tasks then less popular communities. This is a form of preferential attachment, which is often included in models of online communities (Ozmen, Smith, Yilmaz, and Smith 2012, Kumar, Novak, and Tomkins 2010). Communities are assumed to focus on specific topics, and tasks are allocated to communities with similar topics. The topic of a community is not static, but reflects the tasks that a community has recently had success with.

$$P\left(TC_{\{k|TC_k(t)=0\}}(t+1)=x\right)=\left\{\begin{array}{l|l}\frac{1}{\sum_i f(i,k)} & f(x,k)=1\\ 0 & f(x,k)\neq 1\end{array}\right\}$$

$$f(i,k)=\left\{\begin{array}{l|l}1 & TN_k(t+1)\in\left[T_i(t)-\frac{100P_i}{NumUser},T_i(t)+\frac{100P_i}{NumUser}\right]\\ 0 & TN_k(t+1)\notin\left[T_i(t)-\frac{100P_i}{NumUser},T_i(t)+\frac{100P_i}{NumUser}\right]\end{array}\right\}$$

Although the tasks that enter the simulation environment are independent of the communities, which community receives the task does depend on both the size and previous work of the community. Communities are considered to have a topic, bounded in the same domain as tasks, that reflects the specialty of the community. Additionally, larger communities are considered more able to meet a variety of challenges, expanding the range of tasks they can accept. The growth in range is assumed to be linear and such that if all users are in a single community, that community can accept all tasks. The new task generation and allocation occurs once per task completed in the previous round. This way, a population of $NumTask$ tasks is maintained in the simulation environment.

Once a task is assigned to a community, its users are able to contribute to it. In a given community, the users that contribute, and the tasks that get contributions, depend on the number of users with a particular skill and the number of tasks that require that skill. For each community and skill the number of users with that skill, and the number of tasks to which that skill is applicable, are counted. This can be represented as in:

$$NU(i,s)=\sum_{j|UC_j(t)=i,S(j)=s}1$$

$$NT(i,s)=\sum_{k|UT_k(t)=i,TN_k(t+1)\leq s^2,CH_{k,s}(t)=0}1.$$

Only unique contributions are counted, and users that are able to make a unique contribution are assumed to do so. These assumptions mean that for a given community and skill, either every user will contribute or every task will receive a contribution. Of the larger population, a number of members equal to that of the smaller population is chosen to provide or receive the contributions. These assumptions provide probabilities for a contribution occurring for both tasks and users where the probability of a task receiving

a contribution is determined by the relative abundance of users and vice versa.

$$P\left(CH_{\left\{k,s|CH_{k,s}(t)=0,TN+k\leq s^2\right\}}(t+1)=1\right)=\left\{\begin{array}{l|l} min\left(1,\frac{NU(TC_k(t+1),s)}{NT(TC_k(t+1),s)}\right) & NT(TC_k(t+1),s)>0 \\ 0 & NT(TC_k(t+1),s)\leq 0 \end{array}\right\}$$

$$P(A_j(t+1)=1)=min\left(1,\frac{NT(TC_k(t+1),s)}{NU(TC_k(t+1),s)}\right)$$

A trivial consequence of these equations is that for each time-step, in each community, either every user with a particular skill will contribute or every task that requires that skill will have it applied.

Completed tasks are removed from the simulation at the end of every time-step, rather than as they are completed. This can be thought of as time used to verify that the task is complete or to accept a solution. The list of skills applied to a task is used to check if either a user has been able to identify the task as nonprime, or the community has identified a number as prime:

$$Nonprime(k)=\left\{\begin{array}{l|l} 1 & \{s|TN_k(t+1),CH_{k,s}(t+1)=1\}\neq\emptyset \\ 0 & \{s|TN_k(t+1),CH_{k,s}(t+1)=1\}=\emptyset \end{array}\right\}$$

$$Prime(k)=\left\{\begin{array}{l|l} 1 & 1+\sum_s CH_{k,s}>\sqrt{TN_k}(t+1) \\ 0 & 1+\sum_s CH_{k,s}\leq\sqrt{TN_k}(t+1) \end{array}\right\}$$

$$Complete(k)=max(NonPrime(k),Prime(k))$$

Nonprime numbers are detected by checking if any of the contributions to a task divides the task number, if at least one does the number cannot be prime. The identification of prime numbers requires that every skill that is applicable (less than or equal to the square root of the task) has been applied. It is possible for the function identifying primes to give false positives but, as no distinction is made, it does not affect results. Tasks that are completed need to be removed for the model to allow for new tasks. This is accomplished by the following equations:

$$CH_{\{k,s|Complete(k)=1\}}(t+1)=0$$

$$TC_{\{k|Complete(k)=1\}}(t+1)=0$$

The other effect of a completed task is the shift of topic in the communities, which is accomplished with a simple learning algorithm of the form:

$$T_i=(1-\lambda)T_i+\lambda TN_k$$

4.3 Users

The users of online communities are a heterogeneous population searching for something to hold their attention. Users differ in the skill they have for completing tasks, as well as in location. The skill of users is a fixed initial condition so users must move between communities to find tasks to contribute to. Users can leave a community either to join another community or to establish a new community.

Movement of users is driven by an algorithm based on win-stay, lose-shift, as well as preferential attachment (Nowak and Highfield 2011, Ozmen, Smith, Yilmaz, and Smith 2012). Users are assumed to move every round in which they fail to contribute, and where they move to is based on the population of the other communities:

$$P\left(UC_{\{j|A_j=0\}}(t+1)=x\right)=\left\{\begin{array}{l|l} \frac{P_x}{\sum_{i\neq x}P_i} & x\neq UC_j(t) \\ 0 & x=UC_j(t) \end{array}\right\}$$

The other condition for user movement is the foundation of a new community, and it is assumed to increase linearly with the number of abandoned communities:

$$P(UC_j(t+1)=x)=\left\{ \begin{array}{l|l} \frac{rr}{NumUser} & P_x=0 \\ 0 & P_x\neq 0 \end{array} \right\}$$

4.4 Initialization

At initialization users are assigned a skill between two and the square root of *MaxP*, which is the set of possible divisors of task numbers. Communities are assigned a topic from the same distribution as task numbers, and users are assigned to communities. Each of these assignments are assumed to be random.

5 RESULTS

The first characteristic verified is that the population of communities follows a power-law, which can be seen in Figure 1(a). The population of each community is taken as the number of users that contributed to it in the last time step. The method of measurement is meant to capture the active population of communities, which is the only feasible way of measuring population in real data. The distribution is stable in time (not pictured).

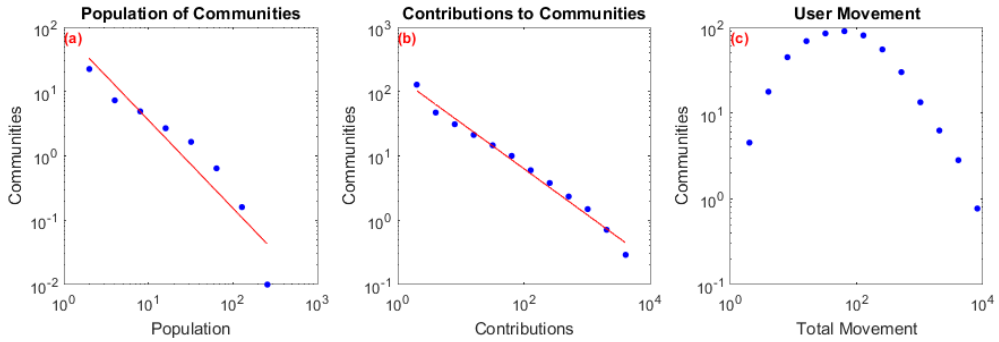

Figure 1: Key distributions of simulated online communities. Plot (a) shows the final distribution of population in communities, as well as the power-law approximation. Plot (b) shows the distribution of contributions throughout the course of the model, as well as the power-law of best fit. Plot (c) shows the distribution of user movement to and from communities over the course of the simulation. Results are averaged for 100 runs with parameters equal to those in Table 2.

Knowing that users are appropriately distributed throughout communities, next checked is that the contributions they make are similarly distributed. Figure 1(b) shows that contributions are distributed according to a power-law, or nearly a power-law. To generate this data, each community tracked the number of contributions it received. Significantly, communities with no users were considered abandoned and had their contribution count reset. This represents the creation of a new community.

Finally, user movement is measured by tallying the total movement in and out of each community. Movement is counted every time a user visits a community regardless of whether or not the user interacts with the community. As can be seen in Figure 1(c), this is clearly not power-law behavior. A possible explanation for this deficit is the difference in considered populations. Results that indicate browsing follows a power-law consider standard browsing behavior. The model was deliberately tailored to model the subset of users that contribute to productive communities. That this sub-population would have abnormal browsing behavior seems plausible, though it has not been confirmed.

Turning now to Stack Exchange, the model will be sampled in the same way as the data, in order to verify that it is capable of producing the same qualitative patterns. Figure 3 shows the result of the model

analysis and should be compared to Figure 2, which shows the Stack Exchange analysis. Details of these analyses follow.

Analysis of the Stack Exchange data yielded four patterns that the model will similarly produce. These patterns include: community size, waiting time for a question to be answered, user movement, and neighbor connectivity. Figure 2 shows these patterns. How the data was sampled to generate those patterns, and what they mean, will be covered in the following paragraphs.

The database of Stack Exchange used to produce the model was downloaded in January of 2014 from (Stack-Exchange 2014). This data set is a freely accessible, anonymized dump of all user-contributed content on the Stack Exchange network provided by Stack Exchange, Inc. under cc-by-sa 3.0 license (Creative-Commons 2014). The data set contained the log files of 110 communities. To give a sense of the data set size, the smallest community—italian.stackexchange.com—was created in November, 2013 and has 374 users, 194 questions, and 387 answers. The largest site, stackoverflow.com (SO) was created in July of 2008 and has 2,728,224 users, 6,474,687 questions and 11,540,788 answers.

Stack Exchange uses a variety of methods to prevent spamming and malicious edits. These methods—including CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) systems, script detection heuristics, new users limits, collective flagging of spam or offensive flags, auto-removal of items based on flags, and human moderators to handle flagged items—work together to form a human-machine combined system that keeps answers clean and effective. Before analyzing the asking and answering activities of users, data was cleaned such that every user who contributed to attention networks had a unique account in the separated log file containing user profile data. This ensures that the activities under investigation were generated by users who had passed the various anti-spam mechanisms of Stack Exchange.

Figure 2 gives a comprehensive description on the population dynamics of the Stack Exchange system. Firstly, it can be observed that the distribution of community size, measured as the population of active users during the period of observation, is highly skewed. The largest community, stackoverflow.com (SO), has more than 2 million users whereas the smallest community has only hundreds of users. In fact, the distribution approximates Zipf's law in two orders of magnitude. Secondly, larger communities are more efficient in solving problems. Figure 2 shows that the average waiting time for accepted answers decreases with community size, indicating that larger communities are more efficient in solving problems. The mobility of users between communities satisfy a "gravity law", which predicts that the number of users moving between two communities is proportional to the product of the "mass" (population) of these two communities. The distribution of community size has a long-tail, i.e., there are only a few very large communities. These communities dominate the user mobility in the entire system. This was confirmed by data in the lower-right panel of Figure 2, displaying the assortativity of the community interaction network. To construct this network, communities are taken as nodes, and the movement of users between these community as edges. For each pair of nodes (communities), the number of moving users, in both directions, is aggregated to obtain weighted, undirected edges. After the network is constructed, edges are removed such that each node maintains its three strongest links, leaving the skeleton from a fully connected network. The constructed community interaction network is disassortative; i.e., high-degree nodes (which are large communities) tend to connect to low-degree nodes (which are small communities). This finding supports the assumption that the mobility of users in the system is dominated by a few large communities.

In short, Stack Exchange is not a very "democratic" system, and is dominated by a few large communities that 1) attract a majority of users in the system; 2) dominate the mobility of users between communities and 3) resolve most of the problems in a very short time. In other words, Stack Exchange activity mostly occurs rapidly in a few large communities, and many small communities rely on the giant communities to provide contributors, most of whom are very likely to return back to the largest communities after they have completed their task.

Since only active users are considered in the data, the model must be similarly restricted. When collecting the data, user activity is determined by posting behavior, that is, the population of a community

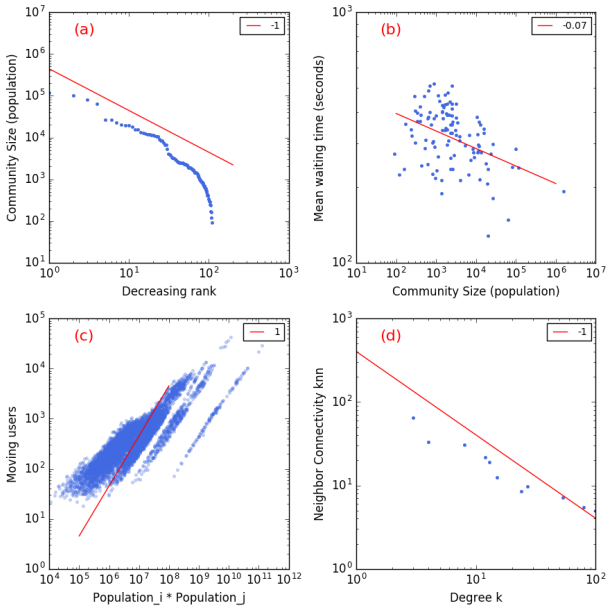

Figure 2: The properties of the communities of Stack Exchange. Starting from the top left plot (a) shows the size of the population of communities against their rank, this is compared to a Zipf law in red plot (b) shows the mean length of time after a question is asked before the community provides an answer plot (c) shows the the undirected movement of users against the product of the community sizes. Plot (d) shows the mean degree of neighbors against a community's own degree, in the network skeleton that preserves the strongest links.

is the number of users that have recently posted. This is replicated in the model by setting the population of a community, at a given time step, to the number of users that made a contribution to that community, during that time step. The population of communities appears to change proportionally with rank for more successful communities, however less successful communities seem to have proportionally larger variation. The model seems to predict more variation in less successful communities than the data would indicate. A possible explanation is Stack Exchange's policy on new communities. To add a community to Stack Exchange it must first prove its viability during a trial period, and no such mechanism exists in the model.

A question posted on Stack Exchange is considered answered only when the user who asked the question has accepted an answer. This can only occur after a minimum waiting period of 15 minutes. Both of these, as well as the discrete time of the model are some of the possible reasons for differences between the model and the data for this metric. The key pattern of waiting time decreasing with community size does hold, which seems to indicate that in both our model and Stack Exchange, the capacity of communities to complete tasks grows more rapidly than the number of tasks they attract.

User movement is tallied over the entire course of the simulation. This is then compared to the population of communities at the final time. This is taken, rather than a continual population count, for simplicity and because community populations stabilize quickly in the model. The result from this analysis is unsurprising and suggests that most user traffic is between large communities with very little between small communities.

Finally, the communities of both Stack Exchange and the model are dis-assortative; that is that the average degree of a community's neighbors is inversely proportional to the community's own degree. The construction of the network is crucial in obtaining this result and identical for the Stack Exchange and model data. Using the data from user movement, each community is linked to the three communities they have the most user traffic with, duplicate links are not counted, and ties are broken randomly. This

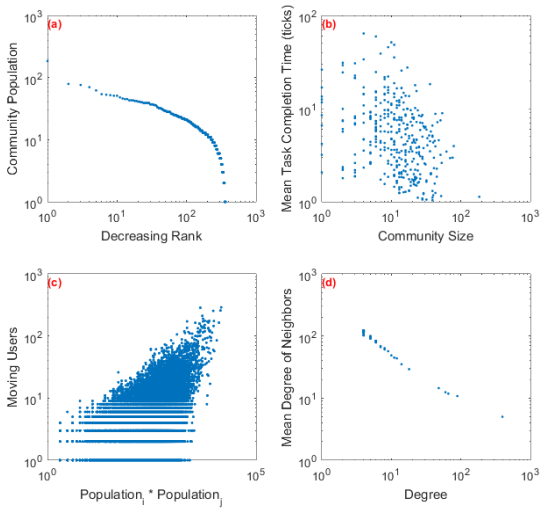

Figure 3: A model of online communities. Starting from the top left plot (a) shows the population of communities against their rank. Plot (b) shows the mean length of time after a question is asked before the community provides an answer. Plot (c) shows the undirected movement of users against the product of the community sizes. Plot (d) shows the mean degree of neighbors against a community's own degree.

results in a network with minimum degree 3, and a maximum degree of one less than the number of communities. On this network, the largest communities are interconnected, as they have high traffic in both directions; however, small communities are linked to large communities rather than being linked to each other. This results in a dis-assortative network where those most highly connected communities have the least connected neighbors.

6 CONCLUSIONS

In this paper, a selection of the initial results of a population ecology oriented model of online communities are presented. Future research will focus on systematic sensitivity analysis to understand how assumptions on which the the underlying mechanisms are based affect the outcomes. Furthermore, we plan to use the model to derive a better understanding of factors that influence the success of communities, and whether application of such a model can be used to predict this same success.

ACKNOWLEDGMENTS

We acknowledge financial support for this work from the National Science Foundation, grant number 1210856 and the Simon A. Levin Mathematical, Computational and Modeling Sciences Center.

REFERENCES

Albert, R., and A.-L. Barabási. 2002. "Statistical Mechanics of Complex Networks". *Reviews of modern physics* 74 (1): 47.

Barabási, A.-L., and R. Albert. 1999. "Emergence of Scaling in Random Networks". *science* 286 (5439): 509–512.

Broder, A., R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. 2000. "Graph Structure in the Web". *Computer networks* 33 (1): 309–320.

Creative-Commons 2014. "Creative Commons License". https://creativecommons.org/licenses/by-sa/3.0/us/. Accessed: 2014-01.

Fu, F., L. Liu, and L. Wang. 2008. "Empirical Analysis of Online Social Networks in the Age of Web 2.0". *Physica A: Statistical Mechanics and its Applications* 387 (2): 675–684.

Huberman, B. A., and L. A. Adamic. 1999. "Internet: Growth Dynamics of the World-Wide Web". *Nature* 401 (6749): 131–131.

Huberman, B. A., P. L. Pirolli, J. E. Pitkow, and R. M. Lukose. 1998. "Strong Regularities in World Wide Web Surfing". *Science* 280 (5360): 95–97.

Huberman, B. A., D. M. Romero, and F. Wu. 2009. "Crowdsourcing, Attention and Productivity". *Journal of Information Science*.

Kittur, A., and R. E. Kraut. 2010. "Beyond Wikipedia: Coordination and Conflict in Online Production Groups". In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, 215–224. ACM.

Kumar, R., J. Novak, and A. Tomkins. 2010. "Structure and Evolution of Online Social Networks". In *Link mining: models, algorithms, and applications*, 337–357. Springer.

Nowak, M., and R. Highfield. 2011. *SuperCooperators: Altruism, Evolution, and Why We Need Each Other to Succeed*. New York: Free Press.

Ozmen, O., J. Smith, L. Yilmaz, and A. E. Smith. 2012. "A Complex Adaptive Model of Information Foraging and Preferential Attachment Dynamics in Global Participatory Science". In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2012 IEEE International Multi-Disciplinary Conference on*, 65–72. IEEE.

Pastor-Satorras, R., A. Vázquez, and A. Vespignani. 2001. "Dynamical and Correlation Properties of the Internet". *Physical review letters* 87 (25): 258701.

Radtke, N. P. 2011. *FLOSSSim: Understanding the Free/Libre Open Source Software (FLOSS) Development Process through Agent-Based Modeling*. Ph. D. thesis, Citeseer.

Stack-Exchange 2014. "Stack Exchange Data Dump". https://archive.org/details/stackexchange. Accessed: 2014-01.

Wilkinson, D. M. 2008. "Strong Regularities in Online Peer Production". In *Proceedings of the 9th ACM conference on Electronic commerce*, 302–309. ACM.

Wu, F., D. M. Wilkinson, and B. A. Huberman. 2009. "Feedback Loops of Attention in Peer Production". In *International Conference on Computational Science and Engineering, 2009. CSE'09.*, Volume 4, 409–415. IEEE.

Yasseri, T., and J. Kertész. 2013. "Value Production in a Collaborative Environment". *Journal of Statistical Physics* 151 (3-4): 414–439.

AUTHOR BIOGRAPHIES

MILES K. MANNING is a graduate student in Applied Mathematics in the Life and Social Sciences at Arizona State University. He holds a Bachelors of Science in Mathematics from Arizona State University. His research interests are in large scale collaboration and population ecology, specifically in how simulation can inform the organization of collaboration. His email address is miles.k.manning5@gmail.com.

MARCO A. JANSSEN is a Professor in the School of Sustainability and Director of the Center for Behavior, Institutions and the Environment, both at Arizona State University. He holds a PhD in Mathematics from Maastricht University, the Netherlands. His research combines computational modeling, behavioral experiments and case study analysis to study collective action problems in human-environmental systems. His email address is marco.janssen@asu.edu.

LINGFEI WU is a post-doc researcher in Knowledge Lab, the Computation Institute at the University of Chicago. He holds a PhD in communication. His research interests are in the dynamics of social attention and the collaborative production of knowledge. His email address is lingfeiw@uchicago.edu.