On the Dynamics of Topic-Based Communities in Online Knowledge-Sharing Networks

Anna Guimarães, Ana Paula Couto da Silva, Jussara M. Almeida
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{anna, ana.coutosilva, jussara}@dcc.ufmg.br

***Abstract*—Online knowledge-sharing networks, such as wikis and Question-Answering (Q&A) portals, offer a rich environment where people can collaborate and share knowledge on specific topics pertaining to their interests and expertise. The continued collaboration of groups of users who share interests in the same topics essentially make up communities, which are here referred to as *topic-based communities*. Analyzing the dynamics of such communities is key to understanding network processes such as the flow of users and information in the network. In this paper, we study how users relate to topic-based communities and how such relationship shapes long-term community dynamics. Using a large dataset collected from Stack Overflow, a popular programming oriented Q&A site, we investigate several factors related to the evolution of communities in the site, such as user participation and the importance of revisits for community sustainability. Moreover, we develop a model to describe community evolution based on user activity which incorporates key aspects of community dynamics, including member revisits and the flow of users across related communities.**

I. Introduction

Online knowledge-sharing networks, such as wikis and question-answering (Q&A) portals, are a unique environment where users with diverse levels of expertise can collaborate to create a specialized knowledge base, participate in discussion threads, seek technical advice, ask questions, and provide answers to fellow users. Due to their premise of serving as alternative sources of information (as opposed to traditional sources, such as mainstream media and literature), much of the existing research on online knowledge-sharing networks has focused on discovering and evaluating quality content [1], [2], finding experts [3], [4], and analyzing audience and contributor profiles [5], [14]. Despite their collaborative nature, and even proven social network characteristics [7], previous research has seldom considered the potential underlying community structure of these networks and the role it plays in its dynamics.

More than a repository of knowledge, online knowledge-sharing networks represent a medium through which users assert their interests in terms of contributions. By posing questions to the community, providing answers with personal expertise, editing and maintaining existing content, users interact with topics of their interest, and with other users who share them. Through their continued collaboration, groups of users who display similar interests essentially make up communities centered around topics of mutual interest and expertise. These *topic-based* communities not only aggregate users based on social interaction [8] but also explicitly associate them with their main interests and contributions in the network.

Analyzing the dynamics of these topic-based communities is key to understanding the flow of users and information in the network, both of which are dictated by how users collectively relate to the topics of discussion they engage in. For example, it is possible to track the success and adoption of a new product or technology by noticing how community dynamics change in response to its introduction as a new topic. Such knowledge could also benefit business services by guiding ad placement decisions or motivating new marketing strategies aimed at a particular community of interest. Similarly, incentive and recommendation mechanisms could be employed by the system to encourage more participation in communities that are losing user interest, or to promote more active communities.

In this work, we study how users relate to these topic-based communities, and how user behavior shapes long-term community dynamics. Because our concept of topic-based communities relies on the network's own structure and on the way members organize their discussions, we are able to move away from the community detection problem [9] and focus on the rich inner dynamics of these communities. Mainly, we are concerned with finding key elements in the relationship between communities and their members, such as member focus and revisits. We also consider the relationship between communities, by recognizing that users display multiple and varied interests, and that their membership in different communities is liable to affect the evolution of all of them.

As a case study, we focus on Stack Overflow[1], a popular Q&A site centered around programming related topics. Our analyses are performed on a large dataset, including almost 20 million posts by 1.7 million users in the 400 most active communities of the site over a six-year period.

Our main contributions are threefold. Firstly, we offer a new perspective on online knowledge-sharing networks, guided by the concept of topic-based communities, that is, groups of users gathered by their shared interest in a particular topic. Secondly, we perform a characterization of a prominent Q&A site in terms of the main topic-based communities it houses. Our analyses uncover key factors in the evolution of communities in the site, and give insight into community aspects like sustainability, user participation, and the flow of users and information in the network. Lastly, building on our characterization results, we develop a new model, called CERIS, to describe community evolution based on user activity. CERIS incorporates key aspects in community dynamics, as identified

[1] http://stackoverflow.com

in our characterization study, such as the impact of revisits and the dynamic flow of users between related communities. Our model expands prior efforts by combining elements of two state-of-the-art approaches [10], [11] to represent the concurrent evolution of multiple communities, and the effect that related communities have on one another.

The rest of this paper is organized as follows. We briefly review related work in Section 2, and introduce the topic-based communities of Stack Overflow and our dataset in Section 3. Our characterization of the dynamic properties of these communities is presented in Section 4, while our community evolution model is introduced in Section 5. We summarize the paper, offering directions for future work in Section 6.

II. Related Work

Research on knowledge-sharing networks primarily focuses on their potential as a rich source of information. In addition to devising methods to uncover quality content and mining expertise [12], [2], [13], [3], recent work seeks also to understand how such expert content comes into fruition. Due to the voluntary and collaborative nature of these sites, a clear approach to the problem is to look at how user behavior in the network translates into meaningful content [14], [7], [5], [6].

Solomon and Wash [15] address possible factors for project survival and sustainability in WikiProjects by relating different measures of project success to three possible growth patterns: accelerating, linear, and descelerating. While we also address factors that impact the evolution of topics in Stack Overflow, our analyses focus on the communities that surround them.

In [16], the authors explore user interaction in Q&A sites by recognizing the existence of communities of users who take part in the same discussions. Complementing previous efforts in community extraction [17], [18], they develop a probabilistic model to extract evolving clusters of linked users from the social graph of Yahoo! Answers and relate these user groups to a shared topic of interest. While they focus on communities formed by user interaction, we here take the communities as given by the topic structure of Stack Overflow and focus rather on their dynamics. Moreover, despite recognizing that users may participate in multiple communities, none of the aforementioned studies analyze how the dynamics of one community may affect the evolution of another, which we do here.

To the best of our knowledge, our work is the first thorough examination of a knowledge-sharing network from the perspective of dynamic topic-based communities.

III. Topic-Based Communities

Collaboration is the building block of a knowledge-sharing network. In wikis, editors cooperate in writing and editing articles, while in Q&A sites, users draw from personal experience and expertise to answer questions from fellow network members, working together to solve specific problems. By collaborating to create a robust knowledge base, users are constantly interacting with topics of their interest, and with other users who share those interests.

In order to describe this multi-faceted relationship, we build on the definitions of social [8] and affiliation networks [19] to introduce the concept of *topic-based communities*. These communities describe groups of users who actively contribute to discussions about given topics of mutual interest. As such, a topic-based community expresses the relationships between collaborating users and the underlying common topic that guides their interactions in the network. Because our definition of topic-based communities derives directly from how members organize their discussions, we can bypass the community detection step in evaluating community structure in a network. For each topic of ongoing discussion in the network, each such community is a well-defined dynamic object.

Various knowledge-sharing networks are available on the Web, each one hosting a variety of topic-based communities. Examples include Quora, Yahoo! Answers[2], and Stack Overflow. As a case study, we focus on Stack Overflow, a currently very popular Q&A network. Next, we introduce Stack Overflow, and then describe the dataset used in our study.

A. Stack Overflow

Stack Overflow is the sub-domain of the Stack Exchange Q&A network specialized in programming questions. Since going online in 2008, the site's primary objective has been to create an open, fully collaborative "library of detailed answers to every question about programming". As of January 2015, the site already hosted over 8 million questions and 15 million answers, and had 3.8 million registered users[3].

In Stack Overflow, the category structure commonly used in Q&A sites is replaced with user-defined tags. That is, instead of posting their questions to one of predefined general categories, users can associate up to five tags to each question. These tags are used to index and organize discussion threads pertaining to the same topics and provide a simple and well-structured way to navigate the website. Some popular tags include "javascript", 'ruby-on-rails-3", and "database".

When applying our concept of topic-based communities to Stack Overflow, each tag corresponds to a topic of interest. Users who have created posts with a certain tag will form the community around that tag (e.g., users who post about the Python programming language form a "Python-based" community). Thus, we take advantage of the structure of the site to intuitively derive topic-based communities.

This definition bears two implications. First, at any moment, a given user may belong to multiple communities as result of the user's participation in multiple discussions about different topics, or even in a single discussion to which multiple tags were assigned. In the latter case, the use of multiple tags suggests that the subject of the discussion relates to multiple disciplines (or topics). Thus, it is reasonable that users involved in such discussions are considered part of all related communities. Secondly, users may use different tags to express the same general subject. We note, however, that Stack Overflow does attempt to eliminate synonyms by periodic moderation of the tags used. Thus, we consider each tag as a different community[4]. However, our definition can be extended to group together multiple related tags in a single community.

[2] http://www.quora.com, http://answers.yahoo.com.

[3] http://stackexchange.com/.

[4] We note that there are no obvious synonyms in the tags used to define the communities in our dataset. Thus, we believe all analyzed communities relate to different topics (in a broader sense).

B. Dataset

The dataset used in our study was built as follows. First, we borrowed the database used in a previous study of answer quality in Stack Overflow [1]. This database contains a complete dump of the system with detailed post data from 2008 to 2012. We then added to the data by collecting more recent content through queries posted to Stack Exchange's Data Explorer[5]. This interface allows the retrieval of the complete database of each sub-domain of Stack Exchange, including Stack Overflow, with a limit of 50 thousand results per query. After a series of queries, designed to cover all posts dated after the prior collection, we were able to extend the original database to a complete dump of Stack Overflow, with all posts from its opening date in August 2008 until August 31^{st} 2014.

We focus our study on the top 400 communities with the largest number of posts (questions and answers), which alone account for over 90% of all posts in the site. Our dataset with the selected communities contains 19.8 million posts made by 1.7 million users over a period of 6 years[6]. Even considering only these 400 communities, we do observe a great variability in terms of number of posts and users across them. On average, each selected community has 100,133 posts (CV[7] of 2.34) and 32,038 users (CV of 1.33)[8]. Futhermore, most communities remained active through a large fraction of the observed period, with an average period of activity (interval between first and last post) of 2,016 days (CV of 0.19).

IV. Temporal Dynamics of Topic-Based Communities in Stack Overflow

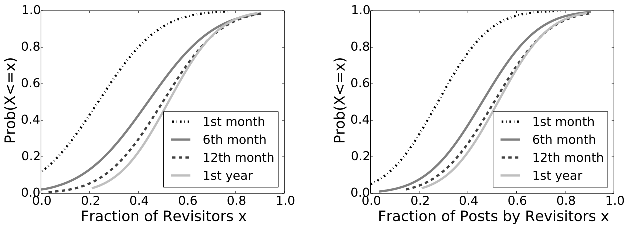
In order to understand the key factors that drive community activity, we look into how users divide their attention across communities in the network. Namely, we focus on the dedication of users to each individual community they participate in and the impact of a shared member base between communities, which result from users exhibiting varied interests.

A. Revisiting Behaviour

We start by analyzing the extent to which users contribute to the same topics (and thus their respective communities) and the effect of this persistent participation on community evolution. Specifically, we investigate users' revisiting behaviour. We define user u as a revisitor of a community c at time t_j if u has made a post in c in any previous instant $t_i < t_j$.

Figure 1(a) shows the cumulative distribution function (CDF) of the fraction of revisitors in all communities over time. Since communities have varied age and member populations, we focus on their first year of activity. We compare three distinct moments in the community's lifetime, namely, the 1^{st}, 6^{th} and 12^{th} months, as well as the overall results in the year, to analyze how member base composition (new users and revisitors) changes over time.

Figure 1(a) shows that, during the 1^{st} month of activity, more than 25% of the members of half of the communities are revisitors, whereas for 25% of the communities, more than 62% of the users are revisitors. These fractions are impressive, given that users only had a short period of contact with the topic in the network. The 6^{th} month of activity shows a leap in revisiting behavior, with the fraction of revisitors exceeding 44% for half of the communities. This fraction continues to grow, albeit at a slower rate, reaching 50% for half of the communities at the 12^{th} month. The distribution is similar if the entire first year is considered: more than 52% of the users in the period are revisitors for half of the communities.



(a) Fraction of revisitors (b) Fraction of revisits

Fig. 1. Revisiting patterns over time (CDFs).

Figure 1(b) shows similar patterns for the fraction of posts by revisitors (i.e., revisits). As members resume participation in previously visited communities, their posts makes up at least 44% of all posts in the community during the 6^{th} month, for half of the communites. The fraction of revisits keeps growing slowly afterwards, falling between 40% and 80% for over 60% of the communities at the 12^{th} month. Thus, despite the great variability, revisitors quickly become a large fraction of the member base of many analyzed communities, also accounting for a large fraction of all monthly posts.

We note that we found no clear correlation[9] between the lifetime of the community (time betwen first and last post) and the fraction of revisitors (ρ = 0.06) as well as the fraction of posts by revisitors (ρ = 0.05). This is not all surprising, as some communities may remain in the system while receiving only few posts, sporadically. We do, however, find a reasonably strong positive correlation between the fraction of revisitors and the total number of posts in a community through its lifetime (ρ = 0.46). Thus, though we cannot claim any causality effect, there is a general trend towards more active communities having higher fractions of revisitors, suggesting that those users play an important role on community sustainability. As an example, in total, the Java community received around 2 million posts, of which 76% were made by revisitors.

B. Participation in Multiple Communities

As the individual expertises and interests of a user may span over various areas of knowledge, a user may participate in any number of communities in the network at any time. Figure 2(a) shows the distribution of the total number of communities a user participates in while in the network. The distribution is highly skewed (CV of 1.54). Around 13% and 15% of the users are involved in only two and three communities, respectively, while the average user participates in a total of 17 communities. Interestingly, the figure shows that only around

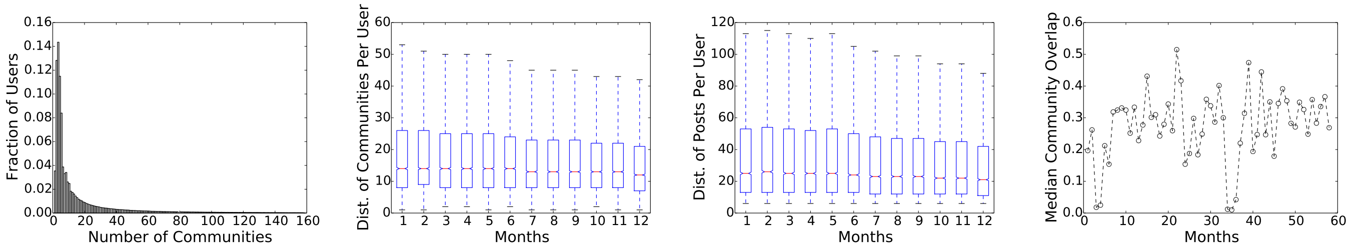
[5] http://data.stackexchange.com/.

[6] Dataset available upon request.

[7] Coefficient of Variation (CV) is the ratio of standard deviation to the mean.

[8] Recall that the same post and user may belong to multiple communities, as a post may receive multiple tags.

[9] We used the Spearman correlation coefficient ρ, as it does *not* require linear relationship between the variables.



(a) Total number of communities (b) Number of communities (c) Number of posts per community (d) Community membership overlap

Fig. 2. User engagement in all communities (aggregated values in a); monthly values in b)-d)).

4% of the users participate in a single community, while 9% of them exceed 50 communities in total.

We also analyze the dynamics of user interests by focusing on the distributions of the number of communities a user participates in on each month, during the user's first year in the network. Figure 2(b) summarizes these distributions by means of boxplots with the 1^{st}, 2^{nd} and 3^{rd} quartiles, as well as the 10^{th} and 90^{th} percentiles. Despite a great variability across users in every month, we note that users, particularly the top-10% that contribute to the largest number of communities (i.e., 90^{th} percentiles), tend to become more focused over time, limiting their posts to slightly fewer topics. The same decaying pattern is also observed for the number of posts by a user in each community, particularly for those who are heavier contributors, as shown in Figure 2(c). Despite this decay, the 10% most active users still go on contributing to at least 42 communities, with at least 87 posts on each of them, even 12 months after joining the network. In contrast, 25% of the users participate in up to 7 communities, with as many as 11 posts to each of them by that time (1^{st} quartile). Thus, in general, users tend to become slightly less active, in terms of numbers of communities and posts, as time progresses, which is consistent with previous results on user activity in Stack Exchange sites (not including Stack Overflow) [5].

We further analyze user interest dynamics by focusing on the specific communities a user contributes to in each window of one month. Given C_t^u, the set of communities a user u contributes to during window t, we employ the Jaccard coefficient[10] J^u to quantify the community membership overlap of u in windows t and $t+1$, defined as $J^u = \frac{\left|C_t^u \cap C_{t+1}^u\right|}{\left|C_t^u \cup C_{t+1}^u\right|}$.

Figure 2(d) shows the evolution of the median Jaccard coefficient, computed across all users, for pairs of consecutive windows. We omit confidence intervals to improve readability, but we note a great variability in these measures (consistent with Figure 2(b)). The plot shows noticeable changes in user interests, with users usually recycling about two thirds of the communities they contribute to and remaining active in the other third. At some points, users come close to joining all new communities, in relation to the previous month. This may be due to the bursty creation of several new communities in the network, which may attract the attention of interested users.

C. Migration of Users Across Communities

As their interests change over time, users may migrate across communities, reducing their participation in some topics to focus on others. Such migration is expected, for example, across communities centered around different versions of the same basic technology (e.g., iOS5 and iOS6)[11].

Figure 3 shows the composition of the member base of two such communities, namely iOS 6 and Ruby on Rails 4, over time. On each month, we split the community members into those who participated in the community centered around the previous version of the technology (i.e., iOS 5 and Ruby on Rails 3), and those who did not. We refer to the latter as *new members*, although they might have participated in communities centered around older versions. As shown in the figure, members inherited from the previous version community are present in the new community in significant number, especially earlier in the lifetime of the new community. For example, at the early stages of the iOS 6 community, one third of its members had been priorly involved in iOS 5 (Figure 3(a)). This fraction is more impressive in Ruby on Rails 4, in which former Rails 3 members make up as much as 82% of the community for the first six months of activity (Figure 3(b)).

This migration of members between topics is not subject only to version-related communities (although they are a clearer example of this phenomenon). In the next section, we show that we can find similar behavior in other related communities with the help of our community evolution model.

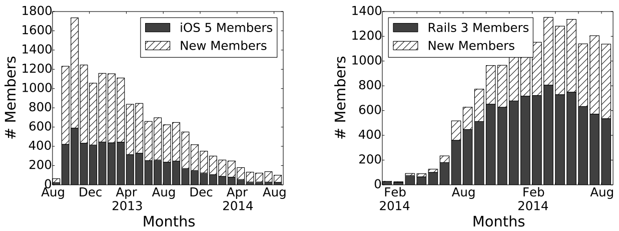
V. Modeling Community Evolution

Our characterization of top communities in Stack Overflow reveals a series of key factors that influence community dynamics. We learned that persisting users are responsible for a significant portion of discussions in many communities, and users often participate in multiple communities throughout most of their sojourn in the network. Thus, communities are not independent objects: a user's membership to one community may impact the dynamics of a different community, if their topics or member base are somehow related. Moreover, despite a great variability in user activity level, we find that all communities seem to be somehow affected by those factors.

Aiming at describing the evolution of communities in the network, we propose **CERIS**, a model that captures the temporal evolution of user activity in a community and the key elements responsible for shaping this activity profile. Unlike the cited models, CERIS captures both revisits by the same user and the interactions between different communities. In the

[10] Statistic commonly used for comparing the similarity of sample sets.

[11] We treat the discussons on each version as a separate community, as different versions of the same technology might have very different features, thus attracting different groups of users.



(a) iOS 6

(b) Ruby on Rails 4

Fig. 3. Member base composition in two example communities.

following, we first present our general modeling approach and describe CERIS. We then show how it fits reasonably well the dynamics of various communities and how it can be used to uncover patterns of user migration across communities.

A. General Approach

The literature is rich with models describing how people share their attention towards a given online social network. Specifically, the community evolution problem is most commonly addressed by adoption models. These models aim at exploring the mechanisms and network conditions that motivate a user's decision to adopt a new technology or join a new community (see [20] and references therein). Both internal (e.g., user interactions and influence) and external factors (e.g., marketing campaigns and word-of-mouth) may be taken into consideration in order to capture the different driving forces behind the evolution of a network of users.

In this context, the author in [20] proposes a reaction-diffusion model that captures the popularity of membership-based websites. Relying on data about daily user activity, the model is able to evaluate the rate of member arrival and departure from the website, as governed by internal factors, as well as the fraction of users targeted by marketing strategies (i.e., potential new members). These parameters assist in fitting the time series data and predicting whether a website will be able to maintain member activity in the future. As its ideas relate closely to our problem, we evaluated this model by applying it to our dataset. Yet, despite its good performance to predict website sustainability, the failed to accurately describe the evolution of communities in Stack Overflow, mainly because it is not able to capture abrupt increases or decay in user activity patterns, nor the interaction among several communities.

Another approach to the community evolution problem is to model communities as contagions, which spread in the network as users become infected by new topics through participation in a community discussion, and eventually recover by ceasing activity in the community. Indeed, Schoenebeck argues that online communities tend to resemble contagious networks, so that applying epidemiology intuition to them should provide a better understanding of their structure [21]. Despite this intuition, epidemic models have more often been employed to capture the dynamics of other types of objects in online network settings, such as information diffusion [22], [23].

Beutel et al. [11] proposed a extension to the classic epidemic SIS (Susceptible-Infected-Susceptible) model to describe the propagation of pairs of infections in a network, aiming at understanding competition effects between two contagions. The authors apply the model to real-world data regarding the adoption of competing products (e.g., video streaming services), leaving other competition scenarios unexplored. Thus, the idea proposed in the paper remains untried in the context of online social communities. In particular, it does not consider the unique aspects which influence how these communities compete, and even cooperate, with one another.

The recently-proposed Phoenix-R model [10] builds on the SIR (Susceptible-Infected-Recovered) model in order to describe the popularity evolution of social media objects (e.g., Youtube videos). A key characteristic of Phoenix-R that distinguishes it from other approaches is the modeling of user *revisits*, by means of a transition into a hidden state from the infected state to imply that users are interacting with the object multiple times. Phoenix-R also captures multiple cascades or outbreaks of interest in the object caused by external events (e.g., a news event about a related subject). Phoenix-R was shown to be robust and outperform other models, like SpikeM [23] and TemporalDynamics [24], in terms of both scalability (to large object collections and long time windows), and accuracy. As Phoenix-R gives special regard to revisits to an object, it seems well-suited for modeling the user activity in our topic-based communities. However, Phoenix-R is restricted to single objects, and thus cannot capture the interaction between related communities and its effect on their activity.

Inspired by the models proposed in [11] and [10], we here propose the *Community Evolution model with Revisits and Inter-community effectS* (CERIS). By combining elements from both approaches, our model is able to not only describe the evolution of a community over time, but also give insight into specific mechanisms that drive community activity, including continued member participation by means of revisits and the impact of related communities on one another.

B. Model Description

For the sake of simplicity, we describe CERIS by focusing on two interacting communities, referred to as C_1 and C_2. Yet, the model is general enough to handle an arbitrary number of communities, at the cost of increased model complexity, as shown in Section V-C. As in Phoenix-R, we assume a fixed population of users who are subject to multiple outbreaks (or *shocks*) of interest in each community. Each shock is modeled as a contagious process directly affecting one given community, although it may also indirectly impact the other one. In the following, we first present the model for a single shock, and then discuss how it generalizes to multiple shocks.

The contagious process happens similarly to a SIS model. Yet, in order to capture the interaction between both communities, we assume users can be either susceptible, infected by C_1, infected by C_2, or infected by both (as in [11]). The recovery from each infection is captured by transitions between these states. Specifically, any user can be in one of 7 states: S, meaning that the user is susceptible to be grabbed to either C_1 or C_2; I_1 and I_2, meaning that user is currently participating in C_1 and C_2, respectively; $I_{1,2}$, meaning that user is participating in both communities,; and V_1, V_2 and $V_{1,2}$, hidden states describing revisits in (only) C_1, C_2 and both, respectively. The total user population (for the shock) is $N = S + I_1 + I_2 + I_{1,2}$. The process evolves as follows:

- At first, an external shock causes interest to arise around one of the communities, say C_1. The shock starts with 1 user infected by the community (I_1=1) and the others susceptible (S=N-1).
- As users keep interacting in the network, new users may join C_1, thus becoming infected by it. This process happens with an infection rate β_1, which determines how contagious C_1 is.
- Users who are discussing one topic may be more frequently exposed to related topics. Thus, infected users in community C_1 may additionally become infected by community C_2 at a modified rate, determined not only by the infectiousness of C_2, i.e. β_2, but also by a measure ε of the relationship between C_1 and C_2. Although ε could be derived by the model (as in [11]), we here estimate its value directly from the input data to reduce computational costs. We estimate ε as equal to the user overlap between both communities, that is, $\varepsilon = \frac{U_1 \cap U_2}{U_1 \cup U_2}$, where U_i is the full set of users who participated in C_i at any point in time.
- The product $\varepsilon\beta_1$ is the rate at which users infected by C_2 are also infected by C_1. Similarly, $\varepsilon\beta_2$ is the rate at which users infected by C_1 become infected by C_2.
- While infected by one or both of these communities, users may continuously interact with them by means of revisits. As in Phoenix-R, we consider that revisits in a community happen as a Poisson process. Parameters ω_1, ω_2 and $\omega_{1,2}$ capture the rates at which users revisit only C_1, only C_2, and both C_1 and C_2.
- Users may eventually cease participating in a community, according to recovery rates γ_1 and γ_2.
- Users who remain active in the network after leaving a community may still come back to it at a later time by a process of reinfection, so that the users may continuously cycle through these states.

Figure 4(a) illustrates these different states and transitions, following a single shock in the network. We assume that the shock starts at time t=0, thus focusing on the dynamics *after* the shock. The following system of continuous-time differential equations describe how the number of users in states I_1, I_2, $I_{1,2}$ and S evolve over time[12]:

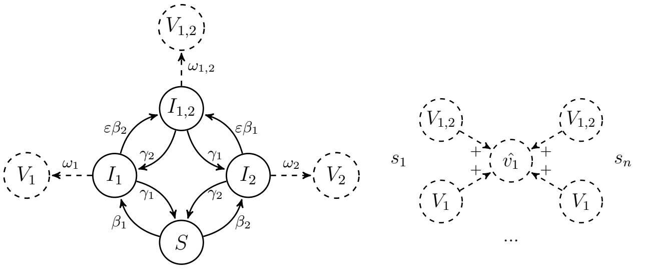
$$\frac{dI_1}{dt} = \beta_1 S(I_1 + I_{1,2}) + \gamma_2 I_{1,2} - \gamma_1 I_1 - \varepsilon\beta_2 I_1(I_2 + I_{1,2}) \quad (1)$$

$$\frac{dI_2}{dt} = \beta_2 S(I_2 + I_{1,2}) + \gamma_1 I_{1,2} - \gamma_2 I_2 - \varepsilon\beta_1 I_2(I_1 + I_{1,2}) \quad (2)$$

$$\frac{dI_{1,2}}{dt} = \varepsilon\beta_1 I_2(I_1 + I_{1,2}) + \epsilon\beta_2 I_1(I_2 + I_{1,2}) - (\gamma_1 + \gamma_2)I_{1,2} \quad (3)$$

$$S(t) = N - (I_1(t) + I_2(t) + I_{1,2}(t)). \quad (4)$$

Equation 1 describes the evolution of the number of users infected by C_1. This process depends on: the rate at which users infected by C_1[13], which is proportional C_1's infectionousness (β_1), are able to influence susceptible users (S); the rate at which users infected by both communities ($I_{1,2}$) leave C_2 remaining active only in C_1, which happens at rate γ_2; the rate at which users infected only by C_1 stop participating in it ($\gamma_1 I_1$); and the rate at which users infected by C_2 (I_2+$I_{1,2}$) infect new users currently participating only in C_1 (I_1), which happens with contagious power $\varepsilon\beta_2$. The latter captures the migration of users from C_1 to C_2. Equation 2 describes the same process for users infected by C_2.



(a) Model states for each shock

(b) Generalization to multiple shocks (two are shown)

Fig. 4. CERIS model.

Equation 3 governs how the number of users infected by both communities evolves. The first two terms capture the rate at which users infected by only one community are infected by the other. The last term captures the rate at which users recover from either community. Finally, Equation 4 describes how the number of susceptible users S evolves over time ($S(t)$) as function of I_1, I_2, $I_{1,2}$, and the fixed population size N.

We note that Equations 1-4 are the same as those proposed in [11] to capture the propagation of pairs of infections in a network. However, unlike in that work, we consider that infected members may repeatedly contribute to the activity of a community by revisiting it. We capture these revisits by hidden states V_1, V_2 and $V_{1,2}$, whose dynamics are defined as:

$$\frac{dV_1}{dt} = \omega_1 I_1, \frac{dV_2}{dt} = \omega_2 I_2, \frac{dV_{1,2}}{dt} = \omega_{1,2} I_{1,2} \quad (5)$$

We can then define the total number of visits (posts) to community C_i at time t as $V_i(t) + V_{1,2}(t)$.

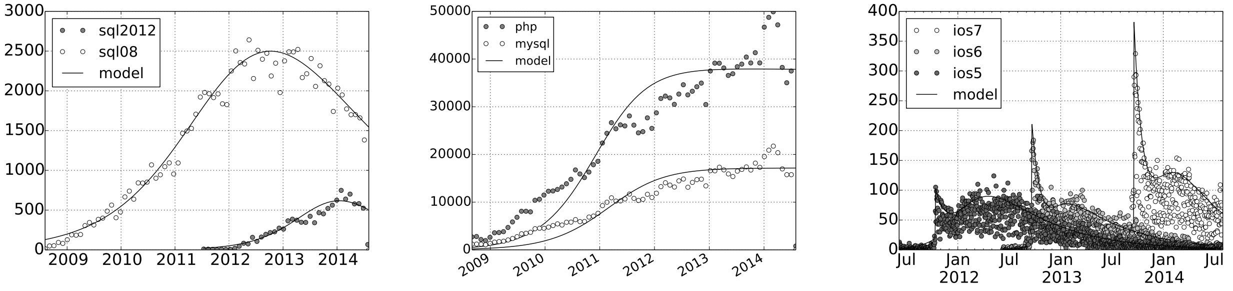
The above description focuses on a single shock. Yet, like Phoenix-R, CERIS also captures multiple shocks that may impact each community. It does so by taking the model illustrated in Figure 4 as a building block for each shock, and connecting the hidden states V_1, V_2 and $V_{1,2}$ so as to aggregate all visits to the same community. This is illustrated in Figure 4(b) for C_1. Note that the connecting point, $\hat{v}_1$, counts the total number of visits in community C_1 due to all shocks.

Specifically, given K the set of all shocks (for both communities), and s_j the time when the j^{th} shock occurred, the total number of visits in community C_i, due to all shocks, at time t is: $\hat{v}_i(t) = \sum_{j=1}^{|K|} V_{i,j}(t - s_j) + V_{1,2,j}(t - s_j)$ (i=1,2).

Moreover, if N_j is the population affected by the j^{th} shock, the overall population of users is given by N=$\sum_{j=1}^{|K|} N_j$. Note that we assume that populations in each shock do not interact with one another. That is, an infected user from shock s_j does not interact with a susceptible one from shock s_p for $j \neq p$. While this may not always hold (users may hear about a topic from different populations), it provides a good approximation, as shown below. It also allows us to have different parameter values for each population, capturing the notion that different populations may behave differently regarding a given topic.

[12] For the sake of simplicity, we use the same notation to refer to both the state and the number of users currently in it.

[13] The total number of users infected by C_1 is given by $I_1 + I_{1,2}$.



(a) SQL Server 2008 and 2012

(b) PHP and MySQL

(c) iOS 5, 6 and 7

Fig. 5. Model fit of the number of posts in related communities.

We now discuss how to fit CERIS to a given dataset representing a set of time series of user activity in each community. We assume time is discretized into time windows (e.g., a month or a day). The fitting procedure is as follows. For each shock j on one of the communities, the model estimates the total number of susceptible users S at time $t = 0$, as well as β_1, β_2, γ_1, γ_2, ω_1, ω_2, $\omega_{1,2}$ from the data, using Equations (1)-(4). These are outputs of the fitting process. To perform the fitting, we follow the approach in [10] to define the set of shocks for each community. The general idea is that each shock corresponds to a peak in the community's time series. We search for peaks by applying a continuous wavelet transform based peak finding algorithm[14], and then apply the minimum description length (MDL) principle [25] to define the total number of shocks $|K|$. By applying MDL, we attempt to find a good tradeoff between model accuracy and model complexity (or generality). In sum, we first find candidate shocks, and then fit the model using the Levenberg-Marquardt (LM) algorithm, adding one shock at a time in decreasing order of peak volume. For each new shock, we evaluate the MDL cost to decide when to stop adding shocks. We refer to [10] for a detailed description of the fitting steps.

C. Model Fittings

We put CERIS to the test by applying it to pairs of communities in Stack Overflow. Because not all communities are expected to be closely related and given the sheer number of possible community pairings ($\binom{400}{2} = 79800$), we limited our tests to all combinations of the top 100 communities, as well as a handful of communities which we knew to be closely related (e.g., iOS 5 and iOS 6). Figure 5 shows examples of the achieved model fitting for daily and monthly activity time series from different sets of related communities.

The fittings were fairly accurate overall, with a mean root mean square error (RMSE) of 21.1317 for all pairs. The model is able to track reasonably well different trends in community activity, including both rise and fall patterns, and multiple peaks of activity. The model is also able to handle related communities and accurately portray their concurrent evolution in the network. We highlight the example in Figure 5(a), where the first signs of activity in the newly created SQL Server 2012 community coincide with a drop in activity in the SQL Server 2008 community. By capturing the migration process of users who go through different stages of community-infection, the model allows us to keep track of the impact one community has on the other. Also, as both communities are evaluated concurrently, the model outputs can be directly applied to compare and contrast activity patterns in different communities, at any given time. For instance, the initial infectiousness β of each community, which stands as a proxy for its attractiveness to new members, provides good insight into how successful a community may grow to be. Indeed, SQL Server 2012 displayed a smaller adoption rate than its predecessor ($\beta_{SQL2008} = 0.00165$ and $\beta_{SQL2012} = 0.00141$), and it never caught up with the SQL Server 2008 popularity, despite having drained a portion of its members.

The model also performs well when analyzing communities which are related but do not display strong migration patterns, such as in Figure 5(b). Instead of competing for members, with users permanently migrating from one community to another, these communities coexist in the network and share a portion of their member bases. We will further discuss the flow of users between communities in the next section. For now, we note how the model is able to handle communities with largely distinct populations and still capture how they evolve both independently, with their own distinct member bases, as well as cooperatively, through their shared members.

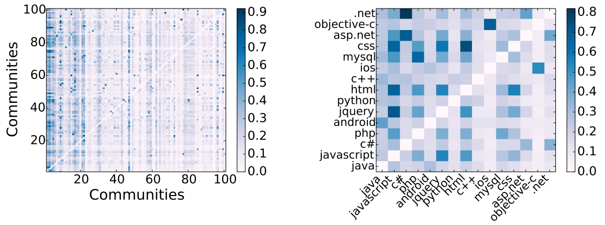
Moreover, as mentioned, CERIS can be easily extended to handle more than two communities. As illustrated in Figure 5(c) for three related communities, the model is able to keep up with changes in community activity patterns, as members transition from one community to another. Thus, although presented as handling pairs of related communities, CERIS can be further extended to account for n-way relationships.

Finally, we note that CERIS is not only reasonably accurate but also quite scalable. The time required to do each of the fittings in Figure 5 is on average only 116 seconds on a Intel Xeon 2.40GHz with 47GB RAM. This is a fairly short time, given the amount and period covered by the data being fitted, particularly in Figure 5(c), which makes use of daily, rather than monthly, activity information.

D. Interactions Between Communities

One key feature of CERIS is its ability to explicitly capture the relationships between communities in the network, based on their shared member base. This is captured by ε, β_1 and β_2, which are outputs of the model. Specifically, we define the *flow* of users from community C_1 to community C_2 as the probability that a user in C_1 will eventually join C_2, which is estimated by our model as $\varepsilon\beta_2$ (similarly, the flow

[14] https://en.wikipedia.org/wiki/Mexican_hat_wavelet



(a) Top 100 communities

(b) Top 15 communities

Fig. 6. Flow of users between communities (source on y-axis, destination on x-axis, color as flow intensity).

in the opposite direction is $\varepsilon\beta_1$). Figure 6(a) illustrates the flow between our top 100 communities as a heatmap, where the color depth of each cell represents flow intensity. Source communities, in order of popularity, are laid out along the y axis, while destination communities are shown on the x axis. The diagonal is left blank as it stands for the flow from a community to itself. This number would be equivalent to the revisit rate, which we have discussed in our characterization.

In general, we find that more popular communities, with high activity levels, have large incoming flows from most communities, including several smaller ones, and their outgoing flow is also distributed among some of these smaller communities. These results are represented by the darker cells for small values of x in Figure 6(a). Thus, popular communities can be seen as hubs in the network, as they accumulate and distribute user activity to different communities in the network.

Figure 6(b) zooms in on the top 15 most active communities in our dataset. Clearly, communities centered around related topics have higher flow values. For example, users in the CSS community have a chance of about 0.72 of participating in the HTML community as well. Interestingly, we often see high flow values in both directions (HTML reciprocates CSS with an outgoing flow of 0.64), indicating that users may transit back and forth between related communities. Nonetheless, more popular communities tend to dominate incoming flow.

The identification of inter-community user migration patterns with CERIS provide valuable knowledge to understand the flow of users and information in the network, supporting further studies on the dynamics of such rich environment.

VI. Conclusions and Future Work

To our knowledge, this is the first deep investigation of the dynamics of topic-based communities in knowledge-sharing networks. Our analyses of these communities in Stack Overflow revealed relevant patterns. We found that, despite great variability, users often participate in multiple communities, frequently changing the community set they belong to. We also found that revisiting users play an important role in community sustainability and that communities are not independent objects, as a user's membership to one community may impact the dynamics of others. These findings drove the design of our CERIS model, which was shown to capture reasonably well the temporal evolution of user activity in a community and the migration of users across related communities.

Future work includes investigating the use of CERIS to predict community activity, and validating our results with other similar systems.

Acknowledgements

This research is funded by the authors' individual grants from CNPq, CAPES, FAPEMIG and the Brazilian National Institute of Science and Technology for Web Research (InWeb).

References

[1] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Exploiting User Feedback to Learn to Rank Answers in Q&A Forums: A Case Study with Stack Overflow," in *Proc. ACM SIGIR*, 2013.

[2] F. Harper, D. Raban, S. Rafaeli, and J. Konstan, "Predictors of Answer Quality in Online Q&A Sites," in *Proc. ACM SIGCHI*, 2008.

[3] S. Ravi, B. Pang, V. Rastogi, and R. Kumar, "Great Question! Question Quality in Community Q&A," in *Proc. ICWSM*, 2014.

[4] A. Pal, R. Farzan, J. Konstan, and R. Kraut, "Early Detection of Potential Experts in Question Answering Communities," in *Proc. Conf. on User Modeling, Adaption and Personalization*, 2011.

[5] A. Furtado, N. Andrade, N. Oliveira, and F. Brasileiro, "Contributor Profiles, Their Dynamics, and Their Importance in Five Q&A Sites," in *Proc. CSCW*, 2013.

[6] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the Social Crowd: an Analysis of Quora," in *Proc. WWW*, 2013.

[7] B. Li, M. Lyu, and I. King, "Communities of Yahoo! Answers and Baidu Zhidao: Complementing or Competing?" in *Proc. IJCNN*, 2012.

[8] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group Formation in Large Social Networks: Membership, Growth, and Evolution," in *Proc. ACM SIGKDD*, 2006.

[9] S. Fortunato, "Community Detection in Graphs," *Physics Report 486.3*, 2010.

[10] F. Figueiredo, J. M. Almeida, Y. Matsubara, B. Ribeiro, and C. Faloutsos, "Revisit Behavior in Social Media: The Phoenix-R Model and Discoveries," *Proc. PKDD*, 2014.

[11] A. Beutel, B. A. Prakash, R. Rosenfeld, and C. Faloutsos, "Interacting Viruses in Networks: Can Both Survive?" in *Proc. ACM SIGKDD*, 2012.

[12] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding High-Quality Content in Social Media," in *Proc. WSDM*, 2008.

[13] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow," in *Proc. ACM SIGKDD*, 2012.

[14] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge Sharing and Yahoo Answers: Everyone Knows Something," in *Proc. WWW*, 2008.

[15] J. Solomon and R. Wash, "Critical Mass of What? Exploring Community Growth in WikiProjects," in *Proc. ICWSM*, 2014.

[16] Z. Zhang, Q. Li, D. Zeng, and H. Gao, "Extracting Evolutionary Communities in Community Question Answering," *JASIST*, 2014.

[17] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng, "Blog community discovery and evolution based on mutual awareness expansion," in *Proc. IEEE/WIC/ACM WI*, 2007.

[18] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, "Community evolution in dynamic multi-mode networks," in *Proc. ACM SIGKDD*, 2008.

[19] E. Zheleva, H. Sharara, and L. Getoor, "Co-evolution of Social and Affiliation Networks," in *Proc. ACM SIGKDD*, 2009.

[20] B. Ribeiro, "Modeling and Predicting the Growth and Death of Membership-based Websites," in *Proc. WWW*, 2014.

[21] G. Schoenebeck, "Potential Networks, Contagious Communities, and Understanding Social Network Structure," in *Proc. WWW*, 2013.

[22] S. A. Myers and J. Leskovec, "Clash of the Contagions: Cooperation and Competition in Information Diffusion," in *Proc. ICDM*, 2012.

[23] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, "Rise and Fall Patterns of Information Diffusion: Model and Implications," in *Proc. ACM SIGKDD*, 2012.

[24] K. Radinsky, K. Svore, S. Dumais, M. Shokouhi, J. Teevan, A. Bocharov, and E. Horvitz, "Behavioral Dynamics on the Web: Learning, Modeling, and Prediction," *ACM TOIS*, vol. 31, no. 3, 2013.

[25] M. H. Hansen and B. Yu, "Model Selection and the Principle of Minimum Description Length," *Journal of the American Statistical Association*, vol. 96, no. 454, 2001.