CARACTERIZAÇÃO E MODELAGEM DA DINÂMICA DE REDES DE COMPARTILHAMENTO DE CONHECIMENTO

ANNA CHRISTINA DE CARVALHO GUIMARÃES

CARACTERIZAÇÃO E MODELAGEM DA DINÂMICA DE REDES DE COMPARTILHAMENTO DE CONHECIMENTO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientadora: Jussara Marques de Almeida
Coorientadora: Ana Paula Couto da Silva

Belo Horizonte

Agosto de 2015

ANNA CHRISTINA DE CARVALHO GUIMARÃES

CHARACTERIZING AND MODELING THE DYNAMICS OF ONLINE KNOWLEDGE-SHARING NETWORKS

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Jussara Marques de Almeida
Co-Advisor: Ana Paula Couto da Silva

Belo Horizonte
August 2015

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Guimarães, Anna Christina de Carvalho

G963c Characterizing and modeling the dynamics of online knowledge-sharing networks / Anna Christina de Carvalho Guimarães. — Belo Horizonte, 2015.
xvi, 53 f. : il. ; 29cm.

Dissertação (Mestrado) - Universidade Federal de Minas Gerais Departamento de Ciência da Computação.

Orientadora: Jussara Marques de Almeida Gonçalves.
Coorientador: Ana Paula Couto da Silva.

1. Computação - Teses. 2. Redes sociais on-line. - Teses. 3. Redes de relações sociais em linha. 4. Web 2.0 – Teses. I. Orientadora. II. Coorientadora. III. Título.

519.6*75(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Characterizing and modeling the dynamics of online knowledge-sharing networks

ANNA CHRISTINA DE CARVALHO GUIMARÃES

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. ANA PAULA COUTO DA SILVA - Coorientadora
Departamento de Ciência da Computação - UFMG

PROF. ARTUR ZIVIANI
Laboratório Nacional de Computação Científica - CNPq

DR. FLAVIO VINICIUS DINIZ DE FIGUEIREDO
Pós-Doc/ DSC - UFCG

PROF. PEDRO OLMO STANCIOLI VAZ DE MELO
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 27 de agosto de 2015.

Resumo

Redes de compartilhamento de conhecimento, como wikis e portais de perguntas e respostas (Q&A) provêem a seus usuários um ambiente para busca e discussão de informação sobre diversos assuntos de seu interesse. Através da sua colaboração para fornecer orientação, levantar e responder perguntas e participar de discussões técnicas, usuários formam comunidades interativas, baseadas em tópicos de interesse mútuo. A análise dessas comunidades centradas em tópicos é crucial no entendimento de processos internos que ditam como a rede evolui com o tempo, os fluxos de usuários entre comunidades e a difusão de informação dentro da rede.

Nesta dissertação, nós investigamos como usuários se relacionam com comunidades centradas em tópicos e como este relacionamento determina a dinâmica das comunidades na rede ao longo do tempo. Utilizando um grande conjunto de dados coletados do Stack Overflow, um site popular de perguntas e respostas sobre programação, estudamos diversos fatores ligados à evolução das comunidades no site, como o impacto de revisitas e da migração de usuários na sustentabilidade de uma comunidade. Nossas descobertas são formalizadas na proposição de um novo modelo de evolução de comunidades que se baseia na atividade de usuários e incorpora elementos-chave da dinâmica de comunidades. Além de descrever os níveis de atividade das comunidades ao longo do tempo, o modelo proposto também permite o entendimento do impacto de comunidades relacionadas umas sobre as outras. Esse conhecimento é então expandido para mostrar como o relacionamento entre comunidades pode ser utilizado para identificar macro-comunidades na rede, com base nos fluxos de usuários.

Palavras-chave: Redes de compartilhamento de conhecimento, dinâmica de comunidades, caracterização e modelagem.

Abstract

Online knowledge-sharing networks, such as wikis and question-answering (Q&A) portals, present users with a channel for seeking and discussing information on diverse subjects pertaining to their interests and expertise. Through their collaborative efforts in providing guidance, raising and answering questions, and otherwise engaging in topical and technical discussions, users form interactive communities around topics of mutual interest. Analyzing the dynamics of these topic-based communities is crucial to understanding the network's inner processes, such as the flow of users across communities and the diffusion of information within it.

In this thesis, we study how users relate to topic-based communities and how this 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, including the impact of user revisits, continued activity and migration on community sustainability. Our findings motivate the development of a community evolution model based on user activity, which incorporates key aspects of community dynamics. In addition to describing the activity levels of communities over time, our new model also provides additional insight into the effects that related communities may have on one another. We expand on this insight to show how information about inter-community relationships can be used to identify macro-communities based on the dynamic flow of users.

Keywords: Knowledge sharing networks, community dynamics, characterization and modeling.

List of Figures

Contents

Chapter 1

Introduction

Knowledge-sharing refers to the exchange of well-understood facts, information and skills previously acquired through experience or education. In a rapidly advancing world where more raw information is divulged than can be effectively processed, it is often difficult to internalize relevant pieces of information, especially in the absence of guidance or a suitable channel through which the information can be conveyed [Cummings, 2003]. In this context, personal knowledge becomes especially valuable and the importance of sharing such knowledge is amplified. When one is faced with a particular problem or simply wants to learn about a given subject, one can then seek knowledge directly from an expert, rather than having to go through several distinct sources to gather possibly loose information that one can then attempt to process.

While knowledge-sharing relies on the interaction between the ones transmitting and the ones receiving it, the development of online technologies has cleared the way for such social interactions taking place among individuals at any time and at any place [Ma, 2012]. As the means and tools for remote discussions become increasingly accessible, online users can more easily seek them out in order to discover, debate and impart knowledge, unimpeded by physical restrictions. Thus, users are limited only by their willingness to engage in knowledge-sharing activities.

On the Web, knowledge-sharing networks exist in various formats, including Wikis (e.g., Wikipedia[1]), discussion forums (e.g., Unix & Linux forums[2]) and question-answering (Q&A) portals (e.g., Stack Overflow[3], Yahoo! Answers[4]). Collectively, these networks represent 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.

[1] http://wikipedia.org/
[2] http://unix.com/
[3] http://stackoverflow.com/
[4] http://answers.yahoo.com/

1.1 Motivation

Focusing on their potential wealth of information, much of the existing research on online knowledge-sharing networks has centered around discovering and evaluating quality content [Dalip et al., 2013; Harper et al., 2008], finding experts [Ravi et al., 2014; Pal et al., 2011] and analyzing audience and contributor profiles [Furtado et al., 2013; Adamic et al., 2008]. Despite their collaborative nature, and even proven social network characteristics [Li et al., 2012], previous research has seldom considered the 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 collectively assert their interests in terms of contributions. By posing questions, providing answers with personal expertise, editing and maintaining existing content, users contribute to topics of their interest and interact 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.

Analyzing the dynamics of these communities plays an integral part in understanding how these networks function and evolve over time. The way users organize themselves in terms of their discussions determines the diffusion of information in the network and reflects their interests in the site at different moments. 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 for discussion. Such knowledge could also benefit business services by guiding advertisement 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 participation in communities that are losing user interest, or to promote more active communities. Insights into the network's overall community structure are also valuable in the design and maintenance of these systems, in order to improve user navigation or to better handle the traffic between specific sections of the network.

1.2 Goals

The primary goal of this thesis is to broaden the understanding of the temporal dynamics of online knowledge-sharing networks. While the results of user collaboration in these networks have been well addressed by previous work, there is still a fundamental gap in understanding the underlying mechanisms that make this collaboration possi-

ble. By investigating the way users organize their efforts and how they continuously relate to different topics of discussion in the network, we seek to explore their potential community structure, which has thus far been overlooked.

Beyond unconvering this community structure, we aim at clarifying user community behavior and how their interactions impact the evolution of topics in the network. Focusing on their behavioral patterns, we wish to discover how users can sustain ongoing topics of discussion over time and how changes in user interests can hinder or benefit a community, in terms of how much attention it receives from other users in the network. Thus, the focus of our research lies not only on a structural interpretation of knowledge-sharing networks, but also in understanding them as dynamic community environments driven by their users.

1.3 Contributions

Our main contributions with this work are threefold.

- **Conceptual.** We shed new light on knowledge-sharing networks by approaching them as a dynamic community environment. To this end, we introduce the concept of *topic-based communities* to describe groups of users gathered by their shared interest in a particular topic or theme. This definition not only captures the social ties between users, which result from their interactions in the network [Fortunato, 2010; Backstrom et al., 2006], but also explicitly relates them to their main interests and contributions.

- **Characterization.** Guided by our new concept of communities in knowledge-sharing networks, we perform a thorough characterization of the prominent Q&A site Stack Overflow in terms of the main topic-based communities it houses. We focus on user behavior in order to understand how it drives community activity over time. Our analyses uncover several key factors in the evolution of communities in the site, and give insight into community aspects like sustainability and continued user participation. By recognizing the diversity of user interests and behavior, we go beyond the inner dynamics of single communities and also consider inter-community user migration patterns as important factors in how communities evolve.

- **Modeling.** Drawing from the main findings in our characterization, we develop the *Community Evolution model with Revisits and Inter-community effectS* (CERIS), an epidemic model that describes the temporal evolution of community

activity. By combining elements of two state-of-the-art approaches to modeling user activity towards a given object [Figueiredo et al., 2014; Beutel et al., 2012], our model is able to represent the concurrent evolution of multiple communities in the network according to key aspects of community dynamics, such as user revisits to a same community and the transition of users from one community to another. By offering a reasonably accurate portrayal of long-term community activity, the model also yields significant results regarding the relationship between communities. In particular, we show how our model can be applied to uncover *macro-communities*, that is, groups of communities related by strong inter-community flows of users, and possibly reflect broader subjects that share a large fraction of experts and/or interested users.

Our work has yielded the following publications:

- On the Dynamics of Topic-Based Communities in Online Knowledge-Sharing Networks, featured in the European Network Intelligence Conference (ENIC) 2015 [Guimarães et al., 2015a], and
- Temporal Analysis of Inter-Community User Flows in Online Knowledge-Sharing Networks, featured in the SIGIR Workshop on Time-aware Information Access (TAIA) 2015 [Guimarães et al., 2015b].

1.4 Outline

The remainder of this work is organized as follows. Chapter 2 reviews existing literature on knowledge-sharing networks and online communities. Chapter 3 introduces the concept of topic-based communities and discusses how it applies to our case study, i.e., Stack Overflow. Our characterization of community dynamics in Stack Overflow follows in Chapter 4. In Chapter 5, we present our community evolution model and discuss its results and applications. Finally, conclusions and directions for future work are presented in Chapter 6.

Chapter 2

Related Work

In this chapter, we review the current literature on knowledge-sharing networks, emphasizing those studies which look into user behavior and evolution, and which are, as such, more closely related to our work. Because the community problem is underaddressed in the context of knowledge-sharing networks, we also discuss relevant studies in community evolution in other social network settings, motivating our community approach. Finally, we present relevant state-of-the-art models which describe user behavior in network evolution and which serve as fundamental stepping stones to our own model of community evolution.

2.1 Knowledge-Sharing Networks

Present research on online knowledge-sharing networks primarily focuses on their potential as a rich source of information. As well as devising methods to uncover quality content and mining expertise [Agichtein et al., 2008; Harper et al., 2008; Anderson et al., 2012; Ravi et al., 2014]), recent work seeks also to understand how such expert content comes into fruition. Due to the voluntary and collaborative nature of these sites, one natural approach to the problem is to look at how users behave in the network and how this behavior translates into meaningful content [Adamic et al., 2008; Li et al., 2012; Furtado et al., 2013; Wang et al., 2013].

Adamic et al. [2008] study user interests on Yahoo! Answers, finding that the frequency and focus of a user's contributions closely relate to the user's personal motivation for using the site. Different goals, such as seeking advice or providing grounded answers to specific problems, will naturally lead users to different categories in the websites and result in different contribution patterns, both in terms of quantity and perceived quality.

In the same vein, Mendes Rodrigues and Milic-Frayling [2009] study the social aspect of Yahoo! Answers and MSN Q&A[1] in order to determine user intent in Q&A networks. Through a manual analysis of a subset of questions from both networks, the authors distinguish social intent in discussions: while some users were mainly interested in factual and practical information, others came to the network looking for opinions, or simply to start an informal conversation, outside of any specific subject. Measurements of the impact of both forms of engagement (non-social and social) reveal that both contribute to the activity in these networks. The authors also suggest that a more detailed analysis of users' social behavior may add to the understanding of their importance in the evolution of the network.

The work of Furtado et al. [2013] describes contributor profiles in Stack Exchange sites by identifying patterns in user activity and behavior, following in the idea that users with distinct intents in knowledge-sharing play distinct parts in shaping the network. The authors analyze these different profiles by grouping users according to their activity levels, their preferred role (i.e., as asker, answerer or commenter), and the quality of their contributions (as evaluated by the number of upvotes their posts have received). Despite finding that only a small minority of users fit into high-activity profiles, the authors argue that both low-activity posters and highly active posters are equally important for content production in the sites. Our work complements this study by investigating user behavior in terms of the topic-based communities they contribute to, giving special attention to the issue of revisiting and sporadic users in community dynamics.

In their study of Quora[2], Wang et al. [2013] investigate how explicit social features (such as the ability to follow other users) affect the perceived quality of questions and answers on the site. As users are faced with an activity feed from the people they follow, posts by popular users, dubbed superusers, who have a greater number of followers will thus have greater exposure. This introduces a bias in how users in the network are directed towards certain discussion threads, so that most of the site's attention will go to a small subset of discussion threads per topic. Although Quora's social features are absent in other Q&A systems, this work suggests the importance of taking user interactions and user influence into account when analyzing site-wide activity.

Zhang et al. [2014] delve further into the interaction between users in Q&A sites by recognizing the existence of communities of users who take part in the same discussions. Focusing on the relationship between askers and answerers, the authors develop a probabilistic model to extract evolving clusters of linked users from the social graph of

[1] http://qna.live.com/. Site closed in 2009.
[2] https://www.quora.com/

Yahoo! Answers and attempt to relate these user groups to a shared topic of interest. Unlike the majority of previous work, which has ultimately been interested in content quality, Zhang et al. approach Yahoo! Answers as a social environment, wherein users interact through discussions of topics pertaining to their interests and expertise. While the focus of this study is mainly on detecting user communities, we here take the communities as given by the structure of Stack Overflow (as we further discuss in Chapter 3), focusing rather on advancing the knowledge on their dynamics. Thus, out work is orthogonal to that of [Zhang et al., 2014]

Solomon and Wash [2014] explore possible factors for project survival and sustainability in WikiProjects. The authors relate different measures of a project's success to three possible growth patterns: accelerating, linear, and descelerating. While the growth of membership is found to correlate with the growth of production (i.e., new articles, edits and reviews), the authors find that a project's sustainability is more closely related to the diversity and interest of its user base than with the content produced. Our work similarly addresses the impact of user participation on the evolution of topics and discussions in Stack Overflow, but unlike [Solomon and Wash, 2014], our main focus lies on the communities that surround these topics. Moreover, despite recognizing that users may have numerous interests in a network, thus dividing their activity across multiple communities, this prior study, like the one by Zhang et al. [2014], does not analyze how the inner dynamics of one community may affect the evolution of another, which we do.

In Zhu et al. [2014], the authors address the issue of overlapping membership across different wikis. This work builds on the idea that communities may benefit from fresh perspectives but suffer from having to share their users' limited attention as they devote their time to many different communities. As wikis are separate websites that require individual membership, the overlapping set of users across them is very small, which makes it difficult to discover significant evidence of the impact of users' shared activity. Despite this, the authors do find a positive impact of member overlap on the survival of communities, particularly in their earlier days when they do not yet have a core set of persisting members. In our work, we analyze groups of users in a same network, where the interaction between different communities is therefore expected to be more significant, thus allowing us to more effectively study the effect of the overlap and migration of users between communities.

The aforementioned studies stand out to us for considering user behavior and dynamics in online knowledge-sharing networks. Despite their different perspectives, all of them point towards factors that may influence content creation and network activity. We consider many of these factors in our own analyses, but focus instead

on community, instead of site-wide, activity. Thus, we take a different approach when studying these networks, while still drawing from previous research results. 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.

2.2 Community Evolution

As previous work demonstrates, user dynamics play a key factor in how a network develops. Furthermore, users do not act in isolation and the set of interactions between them often results in a complex community structure within the network. These communities usually denote groups of densely connected individuals inserted in a larger social network context, such as a circle of friends or a team of peers in a corporation. In our specific context, communities may result from frequent interactions between network participants, such as askers and answerers in a Q&A site [Zhang et al., 2014]. Due to frequent changes in the activity patterns and relationships of individuals, their associated social and communication network is also subject to constant evolution [Palla et al., 2007; Fortunato, 2010]. In this section, we review some current approaches to the community evolution problem in real world and online social settings, and constrast them to our own community approach to Stack Overflow.

Backstrom et al. [2006] study community development and evolution in different types of social networks. The authors rely on data from a co-authorship network, where communities correspond to publication venues and social links correspond to co-authorship relationships, as well as data from LiveJournal, where users can explicitly join pre-established communities in the network and add other users to their friends list. For both of these networks, the authors investigate the structural features that influence the growth of community and its ability to attract new members, as well as how this structure can determine the appearance and diffusion of interests in the network. Key structural features investigated include the total number of members in a community, the number of closed triads (mutual links between three members) and the number of friends a non-member has in the community. The study finds that no single element is capable of dictating community evolution on its own, and instead offers insights into several structural community aspects and methods with which to evaluate them.

Extending the previous work, Leskovec et al. [2008] investigate the evolution of different communities based on link formation between new users and users already in the network. In contrast with other studies, their proposed model takes into considera-

tion the time of appearance and duration of each new link, instead of observing network evolution in different snapshots. With a greater detailing of link dynamics, the authors propose a new method for network evolution based on preferential attachment, which can be used to generate synthetic network data that closely mimics the characteristics of real-world social networks.

In [Alves et al., 2013], the authors study the role of researchers in the evolution of scientific communities. This work hypothesizes that changes in researchers' interests, particularly those researchers who are more well-known, will have an impact on other researchers in the same scientific community. When a research leader decides to leave their current community in order to approach a new theme, they take with them their students, knowledge and resources. This transition may also encourage others to make the same changes, even if they are not linked to the initial researcher. The work observes that there is indeed such an influence of senior members in different scientific communities and that they often feature in the core of the community, bridging together smaller research groups. Comparison of successive moments in the lifetime of these communities also reveals that changes in the community core often lead to changes in the structural properties of the network itself, such as its assortativeness and average degree.

Palla et al. [2007] investigate community extraction and evolution in co-authorship and phone call networks. The authors use a soft-clustering technique to discover cliques of interacting users in the network over several time steps. The resulting communities at each time step are matched with their corresponding communities at the subsequent step, and the evolution of the community is then evaluated according to changes in their member base between time steps, such as the appearance of new members and the departure of old ones. The authors use these results to investigate community sustainability, finding that it strongly correlates with member commitment and that community survival is positively affected by membership turnover, particularly in larger communities, where users are less likely to have a strong relationship with one another.

Similarly, Lin et al. [2007] examine the community structure that arises from mutual awareness between bloggers posting about a specific event or occurrence. In this scenario, communities are formed from the social ties between users who link to each other's blog posts. As in the previous work by Palla et al. [2007], the work presents a method to discover communities from densely connected users at different time steps, which are subsequently matched and compared in order to provide information on community evolution and how relationships between bloggers can change as a function of the events they focus on in their blog posts.

The extraction of evolving communities is also featured in [Tang et al., 2008], which studies communities in multi-mode networks. These networks involve different kinds of participants with different roles interacting at once. A co-authorship network, for example, could be seen as a multi-mode network where researchers can interact with one another as co-authors, with papers as authors and with conferences as committee members, while papers can also interact with conferences via publications. In order to discover communities of the same kinds of participants in these networks, the authors propose a new clustering method which can be applied to a series of network snapshots and regularized to account for community evolution.

While these prior studies deal primarily with the extraction and evolution of social communities, that is, groups of users linked by social interactions, our work focuses on a different community definition, wherein users do not need to be in direct contact in order to be part of the same community. Instead, we rely on topics of interest, which are explicitly declared by users in their contributions in the network, in order to group them together in topic-based communities. This definition, which will be further explained in Chapter 3, moves us away from the community extraction problem, allowing us to focus instead on community dynamics and evolution.

2.3 Ecological Models

The literature is rich with models describing how individuals in a network share their attention and behave towards a given object, be it a topic, an event or a community. Often, this process can occur similarly to well-understood natural processes, such as the adaptation of a living population to an environment or the spread of a viral contagion [Bartholomew and Bartholomew, 1967]. As such, ecological, epidemic and even chemical systems are often adapted into social network settings. We here briefly review some recent studies that apply ecological intuition to diverse situations where individuals interact with one or several objects in the network.

A common approach for describing the dynamics of users' interest in the network is to model the object of their interest as a contagion. The underlying idea is that, similarly to an infectious disease, content can spread from individuals across their social network. By creating and sharing content with those around them, individuals can potentially infect others who will subsequently repeat the process and continue the infection.

In order to describe this contagious process, the classic SIR (Susceptible-Infected-Recovered) model specifies three main stages of infection and accordingly groups indi-

viduals into three states. Initially, individuals who are not yet infected but who have been exposed to an infection are said to be susceptible (S). After coming in direct contact with the infection, individuals may then become infected (I) and are capable of transmiting the infection to other susceptible individuals. Eventually individuals may recover (R) from the infection, becoming imune to it in the future. At any point during an epidemic (that is, while the infection still exists in the network), individuals may be in any of these stages and may transition from one to the next according to an infection and recovery rate. Examples of real-world diseases that can be described by the SIR model include the measles and mumps. A comparable phenomena in a social network setting is the adoption and abandonment of an online social network service. In this case, users who hear about such a service may join and participate in it for some time. Later, they may lose interest, ultimately leading to a permanent departure from the service.

While the SIR model assumes that individuals who recover from a given infection are subsequently immune to it, there are scenarios where an individual may be reinfected multiple times. A biological example of this is the flu, which can affect the same individual several times over. In a social network context, a user may periodically interact with the same content, such as a video or a post repeatedly shared by friends. These situations are addressed by the SIS (Susceptible-Infected-Susceptible) model, where individuals can either be in the Susceptible (S) or Infected state (I). This means that once individuals recover from the infection, they are immediately succeptible to it once more. As in the SIR model, the transitions from one state to another are determined by a rate of infection and recovery, which capture the evolution of the number of individuals in each state during the contagion.

Drawing from these prior models, Matsubara et al. [2012] proposes the Spike-M model to describe and predict rise and fall patterns in information diffusion in different blogging sites. As a parallel to the susceptible and infected states in the previous SIS and SIR models, Spike-M models contagions in two stages: users can either be uninformed about a piece of news or they can be informed and blogging about it. The rate at which users become informed, and thus transition from the first to the second infectious state, is given by the quality or interestingness of the news. Rather than a recovery rate, at which users would transition out of the informed state, the model assumes a decay function that associates the infectiveness of a blogger to the age of the infection. Thus, bloggers become less likely to infect others as time goes by. In particular, the model focuses on the contagion outbreaks, which result in peaks (or spikes) of activity, followed by the decay in activity as users cease to blog or search about a piece of news.

The recently-proposed Phoenix-R model [Figueiredo et al., 2014] also builds on the SIR model in order to describe the popularity evolution of social media objects (e.g., Youtube videos and Twitter hashtags). 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., the release of a new object or a news event about a related subject), offering a more detailed modeling of different aspects driving the popularity of an object.

Myers and Leskovec [2012] also observe contagions in social media. However, instead of focusing on single isolated contagions, this work consider multiple contagions in a same network and the different ways in which they interact, either by cooperating or competing. Contagions may cooperate when two pieces of content are in some way related or similar, so that one may bring attention to the other and thus help the spread of both contagions. On the other hand, in cases where pieces of content are different, they may detract from one another and thus compete for the attention of users. The authors analyze both of these events in the propagation of tweets containing links (e.g.., a news story or video posted elsewhere) and propose a new diffusion model to determine the probability that a user will retweet a post containing a link, based on the content that the user has been previously exposed to (i.e., the sequence of tweets that appeared in their news feed). Thus, rather than focusing on the population of users who take an interest in the subject, the authors focus on how the exposure to different content may benefit or obstruct their diffusion in the network.

Beutel et al. [2012] propose an extension to the SIS model to describe the propagation of pairs of infections in a network, seeking to understand the co-existence and co-evolution of simultaneous contagions. In particular, the model aims to discover a threshold value which determines whether infections can co-exist or whether one will eventually overtake the other. The authors apply the model to both simulated and real-world data regarding the adoption of competing products, such as different video streaming services and internet browsers, finding not only evidence of competition and co-existence but also a possibility for cooperation between infections.

Matsubara et al. [2015] approach the issue of competition and co-evolution in a different light, by comparing products competing for attention to animals competing for food. In this view, the web is considered an ecosystem where user attention is a finite resource and keywords (be them topics or products) behave as living species, which must survive by acquiring these resources. Thus, based on information about user attention directed at certain products on the web (e.g., a search query or a purchase),

the authors develop a model to detect the existence of competition and forecast future dynamics for those products, additionally taking into account non-ecological aspects, such as seasonal events and measurements of keyword interaction.

Ribeiro [2014] proposes a reaction-diffusion model that captures the popularity of membership-based websites, which are potentially competing for members. Relying on data about daily user activity, the model evaluates the rate of member arrival and departure from the website, as governed by internal factors, such as member interactions, as well as the fraction of users targeted by marketing strategies, which aim at attracting new members to the site. These parameters assist in fitting the activity time series of a website and predicting whether it will be able to maintain member activity in the future. This model is further refined in [Ribeiro and Faloutsos, 2015] to more explicitly account for the competition between websites that share a portion of their members and which are, therefore, competing for attention.

The existence of many different approaches to the same basic problem demonstrates that this is an active research topic, which is not yet thoroughly understood. The way users devote their attention to a community or theme may vary according to the specific network setting and its characteristics, and each setting may require a different perspective.

In our work, we make an effort to understand knowledge-sharing networks and take into account their main features in order to develop a model that can accurately portray them. Thus, our model draws from previous modeling approaches and findings, as well as from specific aspects of our case study network and its community structure, which will be evidenced in the following chapters.

Chapter 3

Problem Domain and Contextualization

In this chapter, we formalize our concept of topic-based communities in online knowledge-sharing networks and explain how it differs from the usual community definition for social networks. We then explain how we apply this concept to define communities in Stack Overflow, and present an overview of the website and its structure. Details about the Stack Overflow dataset we use throughout our study are given at the end of the chapter.

3.1 Topic-Based Communities

Collaboration is the building block of a knowledge-sharing network. In wikis, editors cooperate in writing and editing articles, continuously expanding upon each other's work to provide new insights into a same subject. In support forums and Q&A sites, users draw from personal experience and expertise to answer questions posed by fellow network members, working together to solve specific problems. By collaborating to create a robust knowledge base, users are constantly in contact 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 [Wasserman and Faust, 1994] and affiliation networks [Zheleva et al., 2009] 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 defined in the network (i.e., a subject category for an article, a post tag or a keyword). As such, a topic-based community expresses the relationships between collaborating users and the underlying common topic that guides their interactions in the network.

This differs from the traditional definition of social communities in networks, which refers to tightly-knit groups of users within which connections are dense and between which connections are sparser [Girvan and Newman, 2002]. In contrast, because our definition of topic-based communities directly stems from how members organize their discussions, we do not need to derive a community structure from user interactions within the network. For each topic of ongoing discussion, each corresponding community is a well-defined dynamic object.

To further explain these topic-based communities, their inner dynamics and their evolution, we adopt the popular Q&A forum Stack Overflow as a case study. In the following section, we introduce Stack Overflow and explain how we define topic-based in this network.

3.2 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 July 2015, the site hosted over 10 million questions and 16 million answers, and had 4.4 million registered users, over 2 million of which have made at least one post[1]. Figure 3.1 shows currently active questions on the front page of the site.

In Stack Overflow, the category structure commonly used in Q&A sites is replaced with user-defined tags (seen below each of the questions in Figure 3.1). Thus, instead of posting their questions to one of a predefined set of general categories, users can associate up to five key terms, of tags, to their questions, in order to denote the topics being addressed. 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", "database" and "performance".

When applying our concept of topic-based communities to Stack Overflow, each tag corresponds to a topic of interest in the network. Users who have contributed to a certain tag, by creating questions and answers, 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 user-defined structure of the site to intuitively derive topic-based communities.

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

Figure 3.1: Front page of Stack Overflow.

This definition bears two implications. Firstly, 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 and tag redundancy by periodic moderation of the tags used. Thus, we consider each tag as a different community. Nonetheless, our definition can be easily extended to group together multiple related tags in a single community relating to a more general common theme. We demonstrate this with the aid of our community evolution model (Chapter 5) and the discovery of macro-communities based on the interaction of users with multiple topics and communities.

3.3 Dataset

Stack Exchange asserts its aim of maintaining an open knowledge-sharing network by providing nearly irrestricted access to its contents. Data dumps with almost all of

the site's contents are periodically offered for download under a Creative Commons license[2]. Alternatively, the Stack Exchange database can be directly queried in real-time via its Data Explorer[3], allowing access to all data from any of its sub-domains, including Stack Overflow, with a limit of 50 thousand results per query.

Because the data dumps lacked information about post tags (as relationship data is too vast and too unstable to be properly supported in a dump format), we built our dataset by collecting data directly from the site. Initially, we borrowed the database used in a previous study of answer quality in Stack Overflow [Dalip et al., 2013]. This database contains a complete dump of the system with detailed data about posts (i.e. questions and answers) made from 2008 to 2012. We then added to this data by collecting more recent content through queries posted to Stack Exchange's Data Explorer. 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 user and post activity 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[4] with the highest number of posts, which alone account for over 90% of all posts in the site[5]. In total, our dataset with the selected communities contains 19.8 million posts made by 1.7 million users over a period of 6 years. On average, each selected community has 100,133 posts (CV[6] of 2.34) and 32,038 users (CV of 1.33)[7]. 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).

Even a simplified overview of our data demonstrates a great variability in our 400 communities, in terms of the number of posts and uses across them. These and other aspects concerning community and user activity, as well as their evolution, will be expanded upon throughout our characterization of Stack Overflow in the following chapter and will remain our main focus for the rest of the thesis.

[2]https://archive.org/details/stackexchange

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

[4]From here on, we use the terms "tags" and "communities" interchangeably, as each community relates directly to user activity surrounding its corresponding tag.

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

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

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

Chapter 4

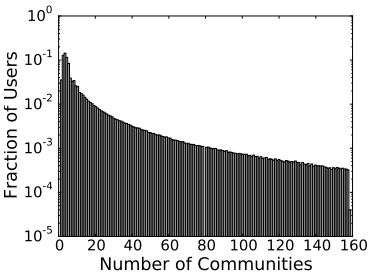
Temporal Dynamics of Topic-Based Communities

In this chapter, we attempt to understand the structure and dynamics of topic-based communities in Stack Overflow. We examine different factors that drive community activity by looking into how users divide their attention across different communities in the network (Section 4.1), how they relate to communities in the long-term (Section 4.2) and how their role in each community affects its sustainability (Section 4.3). Because we are dealing with a multi-community environment, which results from its users exhibiting varied interests, we also look into how communities may affect one another by having a shared member base (Section 4.4). As often as possible, our analyses consider user behavior over time, in order to understand how changes in this behavior may affect community evolution.

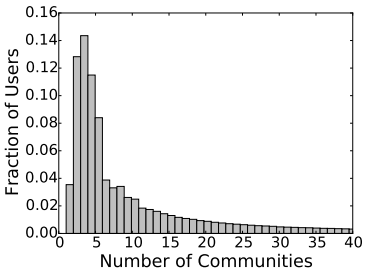
4.1 Participation in Multiple Communities

The individual expertises and interests of a user may span over various areas of knowledge. This implies that, at any time, a user may participate in any number of communities in the network. With that in mind, we start our characterization by studying how users relate to different topics (and, by consequence, their respective communities) in the network, and how their interests change over time.

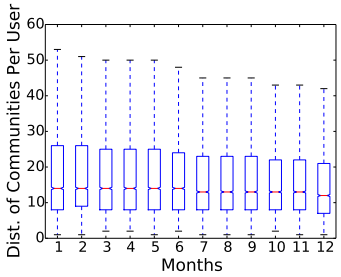
Figure 4.1(a) shows the log-distribution of the total number of communities a user participates in while in the network, with zoomed-in results in Figure 4.1(b). The distribution is highly skewed, with a CV of 1.54. Around 13% and 15% of the users are involved in only two and three communities, respectively, while the average user participates in a total of 17 communities. Interestingly, the figure shows that only



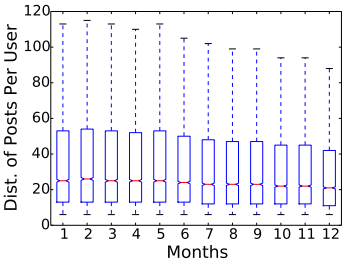
(a) Number of communities (total)



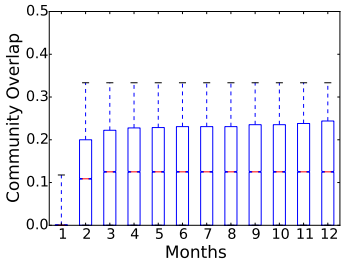
(b) Number of communities (total): zoom on first quarter



(c) Number of communities (over one year)



(d) Number of posts per community (over one year)



(e) Community membership overlap (over one year)

Figure 4.1: User engagement in all communities.

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

Figures 4.1(a) and (b) show aggregated results, considering the whole lifetime of users in the network. However, at one moment, a user may be highly involved in discussions on a certain topic, only to have an entirely different subject steal away their attention in the next. We analyze the dynamics of user interests by focusing first on the number of communities a user participates in on each month, during the user's first year in the network. Figure 4.1(c) summarizes the distributions of the number of communities per user on each month by means of boxplots with the 1^{st}, 2^{nd} and 3^{rd} quartiles, as well as the 10^{th} and 90^{th} percentiles. We note a great variability across users in every time window. The figure also reveals that users, particularly the top-10% users that contribute to the largest number of communities (i.e., 90^{th} percentiles) tend to become more focused in their interests over time, limiting their contributions to slightly fewer topics. The same decaying pattern over time 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 4.1(d). 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 at most only 7 communities, with as many as 11 posts on 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. These results are consistent with previous investigations on user activity in Stack Exchange sites (not including Stack Overflow) [Furtado et al., 2013].

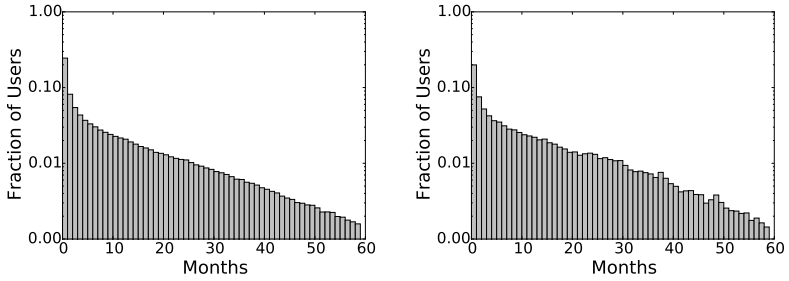
We further analyze user interest dynamics by focusing on the specific communities a user contributes to in each window of one month. We employ the Jaccard coefficient[1] to quantify the community membership overlap in consecutive windows for each user. That is, given C_t^u, the set of communities[2] a user u contributes to during window t, the Jaccard coefficient J^u is 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 4.1(e) shows the evolution of the Jaccard coeffiecient, computed across all users during their first year of activity in the system. The plot shows noticeable changes in user interests, with users consistently reclying nearly 85% of the communities they participate in and remaining active in the other 15%. Note that users may temporarily cease their activity in a community but return to it at a later date, rather than abandon it completely. For instance, it is possible that a user is only ever active in two communities, but alternates between the two in consecutive months, so that our measurement of this user's community membership overlap would always be at zero. Thus, this small overlap in community membership should not be strictly considered a reflection of members' constant (and permanent) change of interests, but rather a portrait of their fluidity in the network and their tendency to jump across different communities, instead of always being restricted to a same group. Indeed, when looking beyond the mean results, we find an overlap of over 0.24 for 25% of members and a maximum overlap of 1.0 (full overlap) in every pair of consecutive time windows. The only exceptions to this behavior occur during the first two time windows, when users are still relatively new to the network and are more likely to be exploring different topics. After this initial period of exploration, similarly to what we find in Figures 4.1(c) and 4.1(d), members begin to return to previously visited communities, slightly restricting their participation in several different ones.

[1] Jaccard coefficient is a statistic used for comparing the similarity and diversity of sample sets.

[2] Since communities are explicitly defined by tags, we can clearly identify those a user contributes to in each window.



(a) All communities (b) Ruby on Rails

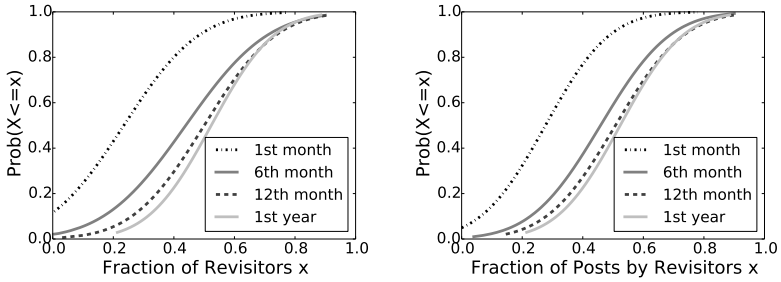
Figure 4.2: Continued activity in a community.

4.2 Continued Activity in a Community

Being so diverse in their interests, users may go through periods of exploration, when they briefly dabble into several subjects, followed by periods of localized activity on fewer topics. We address the way users commit to communities by analyzing how long they remain focused on any given topic.

We express this idea of commitment, or prolonged interest, by quantifying the timespan between a user's first and last post in a *period of continued activity*, that is, a period during which the user has made at least one post per month in a same community. For example, if a member contributed to the HTML community in January, June, July, August, and December, we consider that the user remained active for three periods with durations 1, 3 and 1 months. Thus, we avoid cases of intermitent participation, where a user makes a few posts to a community in one month, but only returns to it several months later, with no activity in between. We note that this does not imply that the user lost interest in the community, as they might have still passively followed the discussion. However, by remaining inactive, the user cannot influence or promote changes in the community evolution.

Figure 4.2(a) shows the log-distribution of the periods of continued activity computed for all users and communities in our dataset. While about 24% of the users stay active in a community for at most a month, over 35% of all users remain continuously active for longer than a year. When relating these results to our previous analysis of how users divide their time across communities in the network, we find an agreement between short periods of continued activity and users dynamic behavior. In particular, the previous Figure 4.1(d) suggested a prevalence of fragmented activity, with users remaining continuously active in only a smaller portion of the communities they once participated in.



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

Figure 4.3: Revisiting patterns over time (CDFs).

Very similar distributions are obtained if we focus on particular communities, as illustrated in Figure 4.2(b) for Ruby on Rails, a popular community in the network. Thus, while communities usually are able to retain some of their members' attention for much longer periods, a large fraction of the members are only (continuously) active for a few months, though they might return to the community, becoming active again, later.

4.3 Revisiting Users

So far, we have learned that despite having varied interests, the average user in Stack Overflow tends to also focus on a subset of topics, sometimes dedicating long periods of continuous activity to them. Now, we attempt to quantify the effect of this persistent participation on community evolution by investigating revisiting behaviour. We define user u as a revisitor of a community at a time window $t{=}i$ if u has contributed to that community in any previous window $t < i$.

Figure 4.3(a) shows the cumulative distribution function (CDF) of the fraction of revisitors in all communities over time. For consistency, since communities have varied ages and member populations, we focus on their first year of activity. We compare three distinct moments in the communities' lifetimes, namely, the 1^{st}, 6^{th} and 12^{th} months of activity, as well as the overall results in all 12 months, to analyze how member base composition (new users and revisitors) changes over time.

Early on, when a topic has only just been introduced in the network, one might expect revisitors to be scarcer. Figure 4.3(a) shows that, during the first month of activity, more than 25% of the users of half of the communities are revisitors, whereas for 10% of the communities, more than 50% of the users are revisitors. These fractions

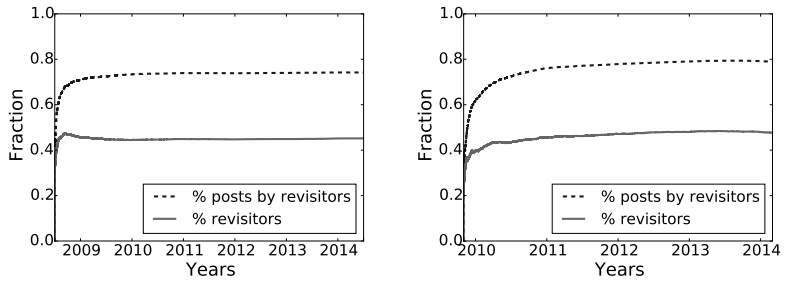
are impressive, given that users only had a short period of contact with the topic in the network. The sixth month of community activity shows a leap in revisiting behavior, with the fraction of revisitors exceeding 44% for half of the communities. This fraction continues to grow as time passes, albeit at a slower rate, reaching 50% for half of the communities at the 12^{th} month. When considering the entire first year, the distribution is similar: the fraction of revisitors in the whole period exceeds 52% for half of the communities.

Figure 4.3(b) shows similar patterns for the fraction of posts by revisitors (i.e., revisits). During their first month of activity, 50% of the communities feature more than 30% of posts made by revisiting users. In time, as members resume participation in previously visited communities, their collective contribution grows to make up at least 48% of all posts in each community during the 6^{th} month, for half of the communites. This number continues to increase slowly afterwards: at the 12^{th} month, the fraction of revisits falls between 40% and 80% for over 60% of the communities. Thus, despite the great variability, we observe that, for many analyzed communities, revisitors quickly become a large fraction of the member base, and account for a large fraction of all monthly posts.

We next analyze to which extent revisiting behavior correlates with community sustainability. First, we found no clear correlation[3] between the lifetime of the community (time between first and last post) and the fraction of revisitors ($\rho = 0.06$) as well as the fraction of posts by revisitors ($\rho = 0.05$). 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 ($\rho = 0.46$). 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, the Java community received around 2 million posts throughout its lifetime, of which 76% were made by revisitors.

We finish this section by illustrating the aforementioned results for specific communities. Figure 4.4 shows the evolution of the fractions of revisitors and revisits during the lifetimes of two particular communities, HTML and Ruby on Rails 3. The results are representative of most communities. After a brief period of activity, the fraction of revisitors to each community catches up to the fraction of new members, and their collective contributions quickly surpass that of newcomers. Past this turning

[3]We used the Spearman correlation coefficient ρ, a nonparametric measure of statistical dependence between two variables that does not require linear relationship between them.



(a) HTML (b) Ruby on Rails 3

Figure 4.4: Temporal evolution of the fraction of revisitors and revisits for two example communities.

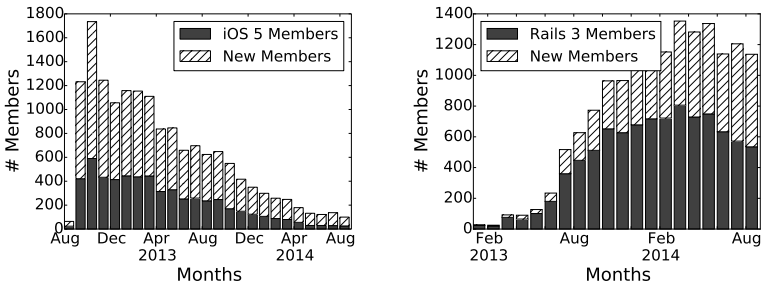
point, the fractions of revisitors and revisits remain roughly stable. Note that, in the long run, even though revisitors account for less than half of the member base (around 45%), these same users are responsible for over 75% of all contributions made in these two communities.

4.4 Community Migration

As user interests change over time, possibly reflecting the rise of a new popular topic (e.g., a new technology), users migrate across communities, reducing their participation in some topics to focus on others of currently greater interest. One particular scenario where this migration is expected is that of communities centered around technologies (e.g., applications, frameworks) that are periodically upgraded to new versions. Topics pertaining to these technologies receive different tags in the network to identify the specific version in question (e.g., iOS 5 and iOS 6), being thus considered different communities[4].

Figure 4.5 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 even older versions (e.g., iOS 4). As shown in the figure, members inherited from the previous version community are indeed present in the new

[4]We treat the discussons on each version separately, as different versions of the same technology might have very different features, thus attracting different groups of users. Additionally, the use of version-specific tags in Stack Overflow is only encouraged when discussing version-specific features.



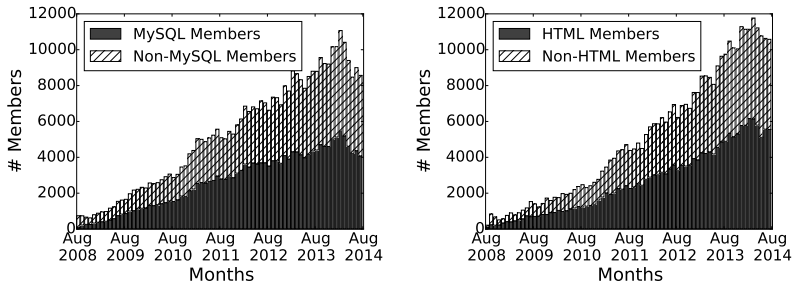
(a) iOS 5 and 6 (b) Ruby on Rails 3 and 4

Figure 4.5: Examples of member composition in communities centered around evolving technologies.

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 4.5(a)). This fraction is more impressive in Ruby on Rails 4, in which former Rails 3 members make up as much as 82% (79% on average) of the community for the first six months of activity (Figure 4.5(b)).

We note, in particular, that the final version of Rails 4 was released in late June 2013, at which point there is leap in community activity. Prior to this, beta versions of Rails 4 were available for download and testing. These were mainly aimed at developers and experienced Rails users, who could provide feedback during the beta process and assist in improving the ultimate final product. Thus, it makes sense that Rails 3 users would be the first to hear about and try out Rails 4, so that they would be the main group discussing the new technology before the official public release. Nevertheless, former Rails 3 members continue to make up the majority of the Rails 4 community, with new members catching up to their numbers only in June 2014, after a year of community activity.

This migration of members between topics is not subject only to version-specific communities, although they are a more intuitive example of this phenomenon. Figure 4.6 shows the member composition of two more example communities, namely MySQl and CSS, in relation to associated communities. Figure 4.6(a) divides the MySQL community between those members who previously participated in the PHP community and those who did not, while Figure 4.6(b) divides the CSS community between previous HTML community members and non-HTML members. While the topics for each community pair are adjacent, as they denote complementing technologies, they are nonetheless independent and discussions about each topic can, in principle, co-exist independently from one another.



(a) PHP Members (b) CSS Members

Figure 4.6: Examples of member composition in communities centered around related technologies.

As in the previous examples, we find a large presence of members from related communities, with as many as 55% of MySQL members having also previously participated in the PHP community and 60% of CSS members who were previous HTML members. These fractions also remain roughly stable throughout the lifetime of the communities, with a mean value of 49.5% (CV of 0.12) for MySQL and PHP, and 51.6% (CV of 0.09) for CSS and HTML, and they seem to usually accompany fluctuations in the total membership of the communities during specific moments (e.g., the increase of members between February and March 2014, shown as the peaks in the Figures). This supports our initial idea that communities do not function and evolve in isolation, based solely on the intra-communitiy activities of their members. Rather, they are also subject to the evolution of related communities, as shared members migrate across communities.

4.5 Summary of Findings

Our characterization of topic-based communities in Stack Overflow uncovers several key aspects of user behavior in terms of how they relate to communities in the network and influence community activity. We find that the average user tends to interact with several different topics, thus participating in multiple communities throughout their sojourn in the network. Users are also not static in their interests, but instead often change the community set they belong to. As a consequence of this variable behavior, we find that communities are not independent objects, but may in fact affect one another as their members migrate across communities, according to changes in their interest or changes around topics. We illustrate this with version-specific topics, where members of a community centered around a previous version of a technology

can make up over 80% of the community centered around a newer version of the same technology. Similar behavior is also found when comparing related topics. For certain cases, related communities were shown to share up to 60% of their member base.

Despite their varied interests and their engagement in multiple communities, users who do go on participating in the same topics are shown to be responsible for a large portion of the community activity surrounding those topics. In over half of the communities we investigate, these revisiting users can account for 52% of the community participants and up to 80% of all contributions made during their first year of activity. This fraction also tends to grow over time, with revisiting users gaining a more prominent role in community sustainability as time goes on.

Overall, we find that users can promote or hinder community activity through their dynamic behavior and their engagement in multiple communities. In the next chapter, we draw from these insights to develop a model for community evolution in the network, which aims to capture both intra-community effects, such as the continued participation of users, and inter-community effects caused by the interaction of a shared member base between communities.

Chapter 5

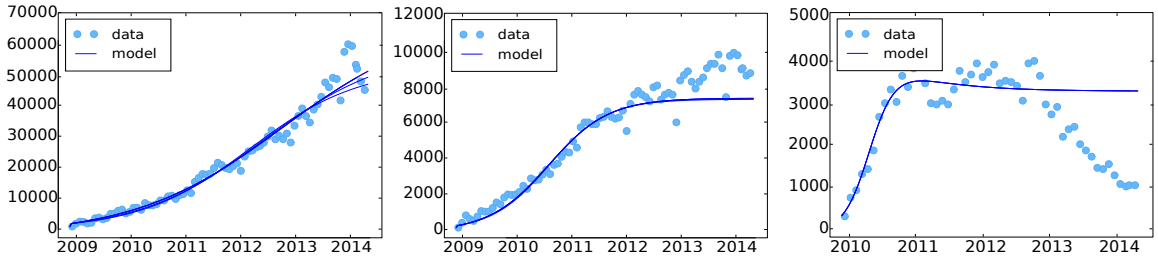
CERIS

Aiming at describing the evolution of topic-based communities, we here 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. The model draws directly from our previous characterization in Chapter 4, explicitly handling community aspects such as revisits by the same users and the interactions between related communities.

In the following sections, we present our general modeling approach (Section 5.1) and describe CERIS (Section 5.2). We then present fitting results, demonstrating it fits reasonably well the dynamics of various communities (Section 5.3). Finally, we discuss applications of the model (Section 5.4) and show how it can be used to uncover patterns of user migration across communities and how communities can be grouped according the flow of users in the network.

5.1 General Approach

In Chapter 2, we presented a variety of models describing how people share their attention towards a specific object. The community evolution problem, in specific, 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 [Ribeiro, 2014]. 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.



(a) HTML (b) Ruby on Rails (c) Ruby on Rails 3

Figure 5.1: Fitting results for the model proposed in [Ribeiro, 2014].

We experimented with some of these approaches in our context, but they failed to capture important elements of community dynamics. As one example, we evaluated the diffusion model proposed by Ribeiro [2014], which describes the popularity evolution of membership-based websites. Although this model was shown to provide good fittings to support predictions of website sustainability, it could not accurately describe the evolution of communities in our dataset, mainly because it relies on adoption thresholds and because the model is not able to capture abrupt increases and decay in user activity patterns, nor the interaction between several communities. Examples of the achieved model fittings for this model are shown in Figure 5.1. Note that the model performs well only as long as there is a steady trend in community activity.

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 [Schoenebeck, 2013]. 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 [Myers and Leskovec, 2012; Matsubara et al., 2012].

We here recall two such models which, while originally designed for other settings, consider similar factors to what we have observed in the dynamics of topic-based communities in Stack Overflow. The first of these models is proposed by Beutel et al. [2012]. The model extends the SIS model to describe competition effects between pairs of contagions in a same network and determine the threshold at which one contagion will overpower another. 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 how they may even cooperate with one another.

A second model of interest is the Phoenix-R model [Figueiredo et al., 2014], which describes the popularity evolution of social media objects (e.g., Youtube videos), with special regard given to revisiting behavior and the cascading behavior of users accessing these objects. Extending the SIR model, Phoenix-R explicitly models user revisits to a same object as a transition from an infected state into an new hidden state, wherein users interact with the object multiple times. Phoenix-R was shown to be robust and outperform previous state-of-the-art models, like SpikeM [Matsubara et al., 2012] and Temporal Dynamics [Radinsky et al., 2013], in terms of both scalability (to large object collections and long time windows) and accuracy. As Phoenix-R explicitly highlights revisits to an object, it seems well-suited for modeling the user activity in our topic-based communities. However, this model 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 [Beutel et al., 2012] and [Figueiredo et al., 2014], 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.

5.2 Model Derivation

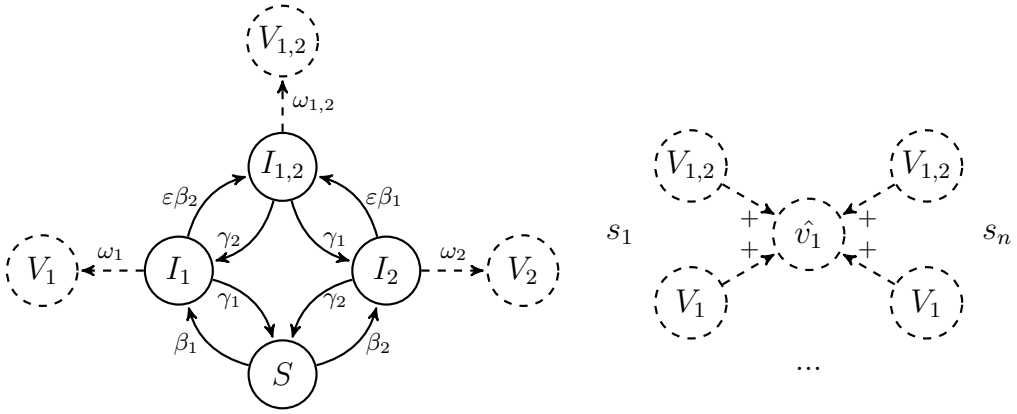
For the sake of simplicity, we describe CERIS by focusing on two interacting communities, referred to as C_1 and C_2. Yet, the model is general enough to handle an arbitrary number of communities, at the cost of increased model complexity, as we will show in Section 5.3. As in Phoenix-R, we assume a 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. 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 Beutel et al. [2012]). 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 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 Beutel et al. [2012]), 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 5.2(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



(a) Model states for each shock

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

Figure 5.2: CERIS model representation.

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[1]:

$$\frac{dI_1}{dt} = \beta_1 S(I_1 + I_{1,2}) + \gamma_2 I_{1,2} - \gamma_1 I_1 - \varepsilon\beta_2 I_1(I_2 + I_{1,2}) \tag{5.1}$$

$$\frac{dI_2}{dt} = \beta_2 S(I_2 + I_{1,2}) + \gamma_1 I_{1,2} - \gamma_2 I_2 - \varepsilon\beta_1 I_2(I_1 + I_{1,2}) \tag{5.2}$$

$$\frac{dI_{1,2}}{dt} = \varepsilon\beta_1 I_2(I_1 + I_{1,2}) + \epsilon\beta_2 I_1(I_2 + I_{1,2}) - (\gamma_1 + \gamma_2) I_{1,2} \tag{5.3}$$

$$S(t) = N - (I_1(t) + I_2(t) + I_{1,2}(t)). \tag{5.4}$$

Equation 5.1 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[2], 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 and remain active only in C_1, which happens at rate γ_2; the rate at which users infected only by C_1 cease 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

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

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

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 5.2 describes the same process for users infected by C_2.

Equation 5.3 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 5.4 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 5.1-5.4 are the same as those proposed in Beutel et al. [2012] to capture the propagation of pairs of infections in a network. However, unlike in that work, we consider the contagious process following one or more shocks in the network and we consider also 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:

$$\frac{dV_1}{dt} = \omega_1 I_1, \frac{dV_2}{dt} = \omega_2 I_2, \frac{dV_{1,2}}{dt} = \omega_{1,2} I_{1,2}. \tag{5.5}$$

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 5.2 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 5.2(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:

$$\hat{v}_i(t) = \sum_{j=1}^{|K|} V_{i,j}(t - s_j) + V_{1,2,j}(t - s_j)(i = 1, 2). \tag{5.6}$$

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 we will show in Section 5.3. 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.

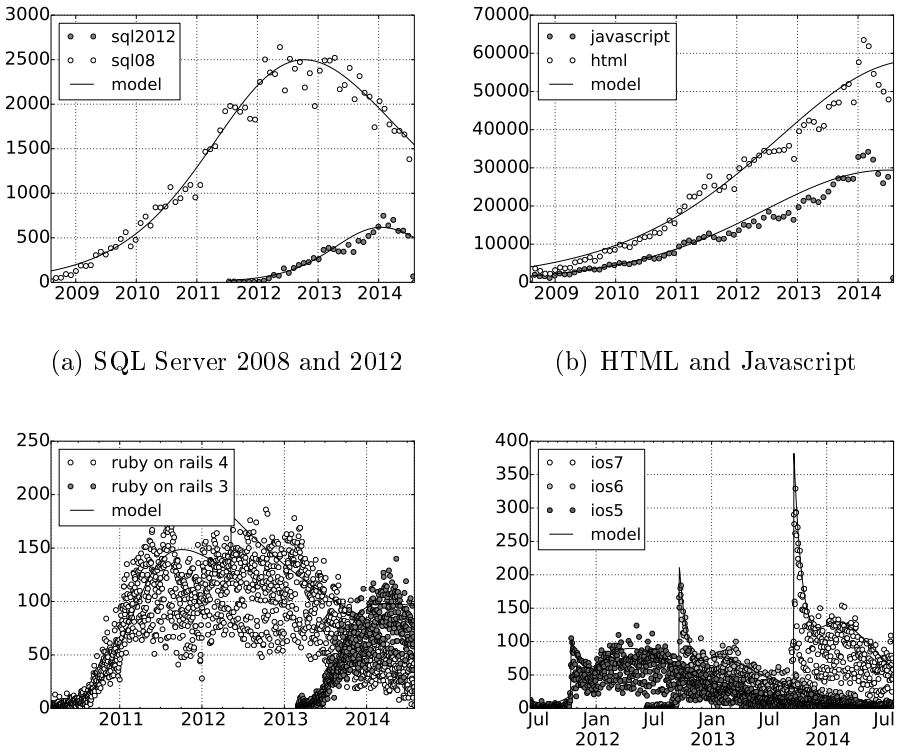
We now discuss how to fit CERIS to a given dataset representing a set of time series of user activity in each community. We assume time is discretized into time windows (e.g., a month or a day). The fitting procedure is as follows. For each shock j on one of the communities, the model estimates the total 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 (5.1)-(5.4). These are outputs of the fitting process. To perform the fitting, we follow the approach in [Figueiredo et al., 2014] to define the set of shocks for each community. Each shock corresponds to a peak in the community's time series. At first, we discover these candidate shocks from activity peaks in the data by applying a continuous wavelet transform-based peak-finding algorithm[3]. We then fit the model using the Levenberg-Marquardt (LM) algorithm, which is a standard approach for nonlinear parameter optimization. The method works by minimizing the sum of the squares of the errors between the data and the model functions [Gavin, 2015]. Shocks are incrementally added, in decreasing order of peak volume. We evaluate the cost and benefit of adding a new shock by applying the Minimum Description Length (MDL) method for model selection, in order to find a good tradeoff between model accuracy and model complexity.

5.3 Fittings

We put CERIS to the test by applying it to sets of communities in Stack Overflow. Figure 5.3 shows examples of the achieved model fitting for monthly (Figures 5.3(a) and (b)) and daily (Figures 5.3(c) and (d))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 and an error below 35.7251 for 75% of all pairs considered. The model gives a reasonably accurate portrayal of the concurrent evolution of related communities in the network and is able to track different trends in community activity, including both rise and fall patterns, and multiple peaks of activity. We high-

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



(a) SQL Server 2008 and 2012

(b) HTML and Javascript

(c) Ruby on Rails 3 and 4

(d) iOS 5, 6 and 7

Figure 5.3: Model fit of the number of posts in related communities.

light the example in Figure 5.3(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. The Ruby on Rails 3 and Ruby on Rails 4 communities behave similarly, as shown in Figure 5.3(c), with the appearance of Rails 4 coinciding with a new shock and subsequent decay in activity in Rails 3. 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 5.3(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 5.3(d) for three related communities, the model is able to keep up with changes in community activity patterns, as members transition from one community to the next. Thus, although presented as handling pairs of related communities, CERIS can be further extended to account for n-way relationships and still provide good results.

Finally, we note that CERIS is not only reasonably accurate but also quite scalable. The fitting process for the model runs at linear time, in relation to the activity time series being fitted. As a concrete example of the model's performance, the time required to do each of the fittings in Figure 5.3 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 Figures 5.3(c) and (d), which make use of daily, rather than monthly, activity information.

5.4 Model Applications

As well as offering a way to describe the evolution of communities in the network, CERIS also give significant insights into the relationship between communities, as dictated by the way users interact with each of them. In this section, we expand on these results and explore some of the applications of the CERIS model and its outputs, including the discovery of migration patterns given by inter-community user flows (Section 5.4.1) and how these can determine macro-communities in the network (Section 5.4.2).

5.4.1 User Flows

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 expressed by ε, which is the measure of the overall overlap between two communities, and β_1

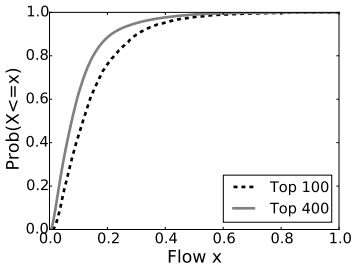


Figure 5.4: Distribution of user flow values between community pairs.

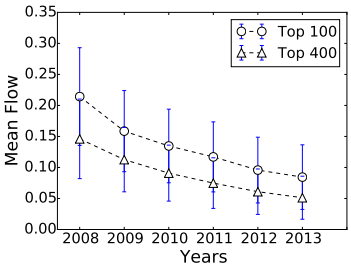


Figure 5.5: Mean flow value and standard deviation over time.

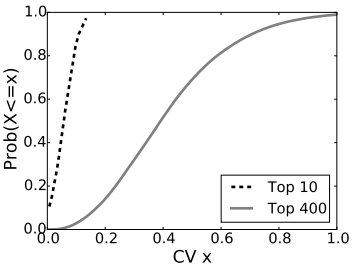


Figure 5.6: Distribution of CV of user flow between community pairs.

and β_2, which are outputs of the model. We utilize these results to 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. In our model, this value is estimated as $\varepsilon\beta_2$ (and, similarly, the flow in the opposite direction is $\varepsilon\beta_1$).

For our following analyses, we look at the flow values output for each monthly time window. Thus, whenever we refer to the user flow between two given communities, we are discussing the average monthly flow of users between them, computed during a certain period of time.

Figure 5.4 shows the cummulative distributions of the computed flow values between all possible community pairs in our dataset, as well as between the subset of the top 100 most active communities, over the whole time period of 2008 to 2014. The distributions are skewed towards lower flow values, especially when considering all 400 top communities, with only 10% of all community pairs displaying a flow value above 0.20. In the smaller subset of 100 communities, we find 25% of community pairs with a flow value of over 0.20 and 50% of pairs with a flow value of at least 0.11. These lower figures are nonetheless still significant: an outgoing flow of 0.11 from community A to community B indicates that members from A approximately have an 11% chance of later participating in community B as well.

We also investigate how these figures change over time by observing the inter-community flows during individual one-year intervals. Figure 5.5 shows the mean flow values and standard deviations computed for community pairs on each year, on both our sets. Over time, we find an increasing number of community pairs with lower flow values. These start at a mean value of 0.21 in 2008-2009 and steadily decrease to 0.08 in 2013-2014. At the same time, the variability (estimated by the coefficient of variation)

of these values increase, with a CV of 0.73 in 2008-2009 and 1.23 at 2013-2014. Thus, over time, community pairs tend to present more distinct relationship levels. This evolution could be partly attributed to the overall popularity growth of the website. As new users join the network with specific intents, they aid in building up distinct community member bases.

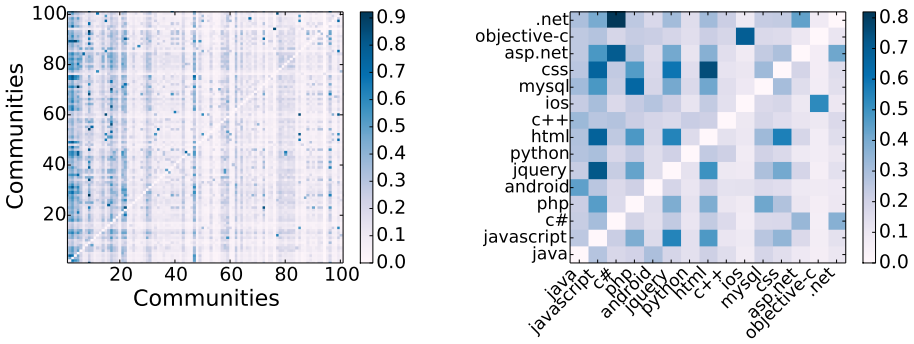
When focusing on specific community pairs, we find diverse evolution patterns of user flow, which are coherent with changes in the relationships between the topics they refer to. As an example, we look at the evolution of the user flow between the Javascript and CSS communities, which correspond to two of the most active topics in the site. In 2008, the outgoing flow from CSS to Javascript was 0.75 and it was met with an incoming flow of 0.55. Six years later, in 2014, the outgoing flow from CSS to Javascript only marginally increased to 0.76, while the flow to CSS from Javascript remained stable at 0.55. Other community pairs, such as Flash and HTML, grew more distant over time, with a flow of 0.43 from Flash to HTML in 2009 and a lower flow value of only 0.29 in 2013.

In some cases, rather than a natural distancing between topics (as with HTML and Flash), we can observe the disruptive effect that one community may have on existing relationships. For example, the Ruby on Rails 3 community starts out with an incoming flow of 0.61 from the Active Record community in 2010. Shortly after the launch of Ruby on Rails 4 and its introduction as a topic in the network in late 2012, this flow value drops to 0.41 in 2013. During this same period, the outgoing flow from Active Record to Ruby on Rails 4 was 0.45. This is a good illustration of how users quickly adapt to the evolution of topics in the network and how the emergence of new technologies (and their respective communities) can impact previously well-established relationships between existing communities.

To summarize these distinct evolution patterns, we estimate the variability in the user flow for each community pair over time by computing the coefficient of variation (CV) of the user flows measured for the pair[4] in all six years covered by our dataset. The higher the CV computed for a given pair (C_1, C_2), the greater variability observed in how the flow from C_1 to C_2 evolved over time. We summarize these results in Figure 5.6, which shows the cumulative distributions of the CV values for all community pairs in our dataset.

Overall, roughly 70% of all pairs had a CV below 0.5 and less than 1% had a CV over 1.0. Thus, most inter-community flows in the network tend to remain roughly stable over time, suffering from moderate to little variation in consecutive

[4]Each pair appears twice, once for each direction (C_1, C_2) and (C_2, C_1).



(a) Top 100 communities

(b) Top 15 communities

Figure 5.7: Flow of users between communities (source on y-axis, destination on x-axis, color as flow intensity).

time windows. As a more pronounced example of this, we also single out the top 10 communities which presented the highest flow values in 2008, also shown in Figure 5.6. Only two of these community pairs had a CV above 0.1, an already low value in itself. We note that this set of communities mainly refer to broader topics (such as "Ruby", as opposed to "Ruby on Rails 3") and therefore may be more robust to temporal changes in the network.

Finally, Figure 5.7(a) illustrates the average monthly 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 (Chapter 4).

In general, we find that more popular communities, with high overall levels of activity, have large incoming flows from most considered communities, including several smaller ones. Yet, 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 5.7(a). Thus, these popular communities can be seen as hubs in the network, as they accumulate and distribute user activity throughout different communities in the network.

Figure 5.7(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 direc-

tions (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.

5.4.2 Macro-Communities

As communities and their relationships become more well-established in the network, it is possible to find patterns between them, such as a common theme. Thus, by observing users' behavior and trajectory in the network, we can recognize groups of related communities purely based on inter-community flow dynamics, rather than having to rely on user interactions or semantic similarity [Papadopoulos et al., 2012].

In order to identify these macro-communities, we approach the network as a connected graph, wherein each node represents a community and edges represent the flow of users between communities. We then apply the Clique Percolation Method (CPM)[5] [Derényi et al., 2005] over the community graph in order to discover clusters of related communities. The method works by discovering k-cliques, that is, fully connected subgraphs of k nodes. These cliques are then joined with adjacent cliques if they share $k-1$ nodes. The resulting clusters therefore correspond to the maximal k-clique-connected sub-graph.

When applying the CPM, we consider only the top 10% of edges with the highest flow values, which denote a more significant relationship between the linked communities. The CPM additionally discards a small number of communities which only appear in isolated small cliques (with 3 nodes or less). This corresponds to 13% of nodes in our top-100 set and 9% of nodes in the top-400 set. We also vary the clique size k, starting at $k = 4$. Exluding very low ($k < 6$) and high values ($k > 15$), results were consistent for different k values[6], with varying k sizes yielding the same number of clusters and similar community sets for each cluster. Thus, for consistency and generality, we fix $k = 6$ for the following analyses.

Figure 5.8 illustrates the macro-communities we found on the top-100 set, considering an aggregated view of flows over the whole 6-year period of our dataset. Overall, we find a relatively small number of community clusters, with at most 5 clusters in the top-400 community set and 3 in the top-100 set. Both cases feature one larger cluster, containing over half of the communities (259 in the top-400 set and 51 in the top-100 set). These clusters include very popular communities (with the greatest number of posts), to which several smaller "satellite" communities are connected with high incom-

[5]Implementation available at http://github.com/michelboaventura/rcpm
[6]Lower and higher values of k often resulted in a single cluster.

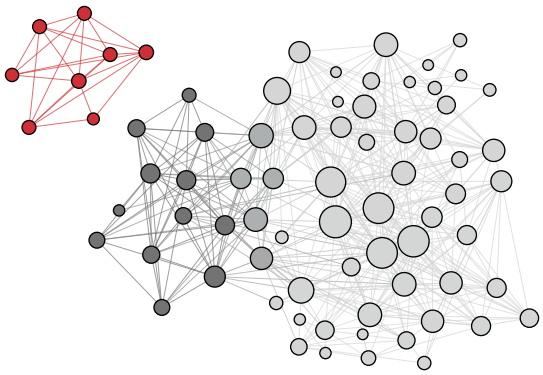
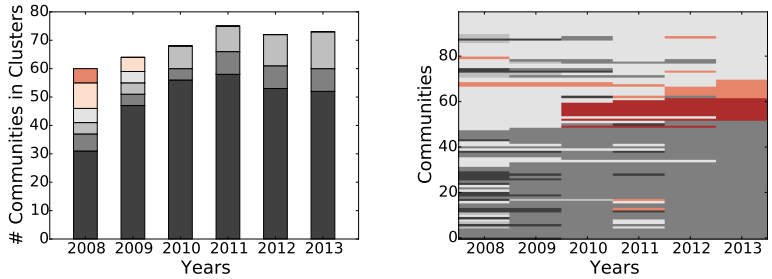


Figure 5.8: Macro-Communities in the top-100 set.

ing flows. As an example, in the top-100 subset, the "Javascript" community was still connected to 88 communities after the removal of the 90% edges with the lowest flow values. This again points towards popular communities acting as hubs in the network, as they gather and redistribute users from and to smaller communities.

These popular communities have another interesting effect in macro-community composition. As macro-communities describe groups of densely connected communities, we expect the average flow between communities in a same cluster to be greater than the flow between distinct clusters. However, because the CPM allows communities to simultaneously belong to different clusters, exceptions do occur. In both our sets, two distinct clusters featured the same four very popular communities (namely, those surrounding the ".net", "c#", "windows", and "asp.net" tags). This makes it so that their overall shared member base is similar, which results in a high flow across clusters. In the top-100 set, we found both an incoming and outgoing flow of 0.31 between the two macro-communities, which stands above the average flow of 0.28 across different macro-communities.

The macro-communities we discovered in our aggregate analysis are nonetheless internally cohesive, both in terms of the user flow across their communities and in terms of their underlying common topic. Among the three macro-communities in our top-100 set, one seems related to general programming discussions, including a majority of communities surrounding programming languages and operating systems. Another cluster refers to Windows and technologies commonly associated with it, such as Visual Basic and the .net framework. The third cluster, which interestingly featured no community intersection with the other two clusters, encompasses discussions strictly related to Apple products and associated technologies, such as iPad, iPhone, OSX



(a) Number of communities per cluster (b) Community membership in clusters

Figure 5.9: Temporal evolution of macro-communities.

and Objective-C. In addition to these three, the top-400 set also features one macro-community surrounding programming IDEs and their many extensions, such as Eclipse, NetBeans and jUnit, and another macro-community exclusively about Ruby and Ruby-on-Rails technologies.

In an effort to better understand the dynamic relationships between the communities in the network, we again analyze their evolution in individual one-year periods. Figure 5.9(a) illustrates the clusters/macro-communities we identified in the top-100 set in each year, along with the number of communities that belonged to each cluster in that year (each bar is divided into a number of sectors corresponding to the number of clusters identified in the year). As in our aggregate analysis, every year sees the presence of one large cluster, involving a majority of the communities, accompanied by a few smaller clusters. These are more fragmented during the early years of the network and display varied compositions. For instance, 6 macro-communities were identified in 2008. In later years, as communities and their relationships grow more established, these clusters also become more distinguished (converging to 3 clusters in 2010) and begin to feature a similar core of communities over time. The larger cluster continuously features a same group of core communities (the same 24 communities were present in the cluster in every year) while also drawing in new communities. As new topics appear in the network over time, communities arise that may help bridge existing communities, thus influencing the make-up of clusters at different periods.

This variability in cluster composition is displayed in Figure 5.9(b), which relates each community to its containing cluster (a different color is used to represent each cluster[7]) in each year. We see how individual communities can belong to different macro-communities in different periods, which translates how inter-community

[7]Light-gray is used for communities that did not belong to any cluster.

flow dynamics vary over time. Nonetheless, we see a tendency towards communities settling in to specific clusters as time goes on, which points to their growing maturity as topics in the network. The few communities which continue to feature in different clusters are often those that show up in the overlap between macro-communities and that thus relate to multiple disciplines (e.g., discussions about Microsoft's C# programming language may appear in the programming macro-community or in the smaller Windows-based macro-community).

While we were easily able to relate these macro-communities to a common general theme, we emphasize that we relied solely on the flow of users for the discovery of macro-communities and no evaluation of the discussions in the associated communities was necessary. Discovering macro-communities in this way is therefore a novel form of community detection, as it relies neither on textual attributes to determine semantic similarity between topics in the network, nor on the social graph of user interaction [Papadopoulos et al., 2012], which may be an ineffective portrait of a community when there are hundreds of thousands of active members posting about the same topics but who have a small chance of interacting directly with one another. Instead, our approach focuses on the relationships between users and the varied topics and communities they engage in over time.

5.4.3 Summary of Findings

In an effort to summarize community dynamics in Stack Overflow, we design a new model to describe the evolution of community activity, while explicitly capturing the effects of user revisits to a single community and the impact of community activity in related communities, which may hinder or contribute to the evolution of one another. In particular, insights about inter-community relationships, which are quantified with the help of the CERIS model outputs, give way to a series of further analyses concerning user activity and behavior.

By focusing on how users interact with multiple communities in the network over time, we establish and investigate the flow of users across communities as a measure of their relationship. This allows us to uncover significant patterns in how users traverse communities and how changes in these patterns can affect the community network. We found that the evolution of user flows tend to follow changes in the technologies themselves, with flow values decreasing between their corresponding communities as technologies drift apart. Nonetheless, most relationships tend to remain stable. Notably, communities regarding more general and more popular topics suffer only from very little variation in their user flows over time.

The most popular communities, with the highest number of users and posts, play another interesting part in the network. As they are connected to a majority of the other communities with high incoming user flows, these communities can potentially act as hubs in the network and bridge communities that would otherwise be disjoint.

We further explore the idea of inter-community user flows to show that it can be used to identify groups of closely related communities in the network. Employing a community detection method, we discover five such macro-communities in the network, each one relating to a broader common topic. The flow of users in the network thus provides a clear portrait of a larger topical structure in the network which can be inferred even without any textual analysis of posts and topics themselves.

These discoveries motivate the use of CERIS for more than a description of community evolution and of the elements that drive it, and exemplify some of the possible applications for its results on inter-community dynamics. Furthermore, these new analyses support our initial approach to the network by providing evidence on its cohesive community structure, which derive from users' participation in different topics.

Chapter 6

Conclusions and Future Work

In this thesis, we investigated knowledge-sharing networks from a novel perspective, by approaching them as a dynamic multi-community environment whose structure is given by the way users organize their discussions in specific topics. Our work therefore complements previous studies, which have mostly focused on exploring the content produced within these networks.

Our community approach to online knowledge-sharing networks was guided by our concept of topic-based communities, which denote groups of users who contribute to specific topics in the network. In order to explain these communities, we focused on data from Stack Overflow, a prominent Q&A site. We performed a thorough characterization of user behavior in the network, investigating how they relate to different topics in the network and how their interests change over time, affecting their activity in different communities that may be part of. Our characterization revealed that, in particular, community activity is driven by two key factors: intra-community dynamics, such as and persisting membership and revisiting behavior, and inter-community dynamics, wherein individual communities that share a portion of their members are able to affect one another.

The findings in our characterization drove the design of a new community evolution model, CERIS. Our model builds on state-of-the-art approaches and incorporates key elements of community dynamics in order to describe the temporal evolution of user activity in topic-based communities. The model was shown to perform reasonably well, yielding overall low fitting errors and a good portrayal of the concurrent evolution of related communities. In addition, the model results give valuable information about community relationships in the network. In particular, we focused on the flow of users between communities as a measure of how users from one community transition into another.

Possible directions for future work include bridging the gap between our study in user behavior and community evolution and previous studies in content and service quality in online knowledge-sharing networks. Investigating how community dynamics affect the perceived quality of the network could be a valuable tool for the improvement of these systems. By exploring repeating patterns of user activity, such as activity peaks or seasonality, and relating these patterns to the content produced at specific times, we may be able to anticipate the production of similar content in the future. For instance, at the early stages of a community based around a newly-created technology, we may find a larger number of discussions regarding its basic functions, while more in-depth discussions may follow later on, as users become familiar with the technology. As similar technologies arise and appear in the network as new topics, displaying similar patterns of activity, the system may then present users with a pre-made set of basic discussions, thus making the topics more easily accessible and facilitating the assimilation of the new technologies.

While we have delved into many of the main factors behind community activity in Stack Overflow for our characterization, there are other site-specific features that may also contribute to our understanding of the site and how its members behave within it. Aspects such as the influence of user badges, status and moderation feedback may influence how users behave and what topics they choose to focus on. Built-in recommendation mechanisms such as the existing "Related" section, which links users to questions similar to the one currently being viewed, also help shape how users navigate the site and discover new discussion threads and topics. Analyzing the impact of these elements, as well as the way users traverse different topic in the site, would thus be an interesting step to further our present study of Stack Overflow.

Another direction would be extending the model to cover topic-based communities found in different systems, such as general-purpose online social networks and discussion boards. As our model incorporates key dynamics of user behavior in a specific knowledge-sharing network, the model may need to be readjusted to account for other such factors when applied to a different setting. Comparing these differences is also an interesting research possibility, as it may aid in understanding the distinct patterns of evolution and sustainability of different networks.

Bibliography

Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: Everyone knows something. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 665--674, New York, NY, USA. ACM.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 183--194. ACM.

Alves, B. L., Benevenuto, F., and Laender, A. H. (2013). The role of research leaders on the evolution of scientific communities. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 649--656. International World Wide Web Conferences Steering Committee.

Anderson, A., Huttenlocher, D., Kleinberg, J., and Leskovec, J. (2012). Discovering value from community activity on focused question answering sites: A case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 850--858, New York, NY, USA. ACM.

Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44--54. ACM.

Bartholomew, D. J. and Bartholomew, D. J. (1967). *Stochastic models for social processes*. Wiley London.

Beutel, A., Prakash, B. A., Rosenfeld, R., and Faloutsos, C. (2012). Interacting viruses in networks: can both survive? In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426--434. ACM.

Cummings, J. (2003). Knowledge sharing: A review of the literature.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2013). Exploiting user feedback to learn to rank answers in q&a forums: A case study with stack overflow. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 543--552, New York, NY, USA. ACM.

Derényi, I., Palla, G., and Vicsek, T. (2005). Clique percolation in random networks. *Physical review letters*.

Figueiredo, F., Almeida, J. M., Matsubara, Y., Ribeiro, B., and Faloutsos, C. (2014). Revisit behavior in social media: The phoenix-r model and discoveries. *arXiv preprint arXiv:1405.1459*.

Fortunato, S. (2010). Community detection in graphs. *Physics Report 486.3*, 486:75–174.

Furtado, A., Andrade, N., Oliveira, N., and Brasileiro, F. (2013). Contributor profiles, their dynamics, and their importance in five q&a sites. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1237--1252, New York, NY, USA. ACM.

Gavin, H. P. (2015). The levenberg-marquardt method for nonlinear least squares curve-fitting problems. In *Department of Civil and Environmental Engineering, Duke University*.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.

Guimarães, A., Silva, A. P. C., and Almeida, J. M. (2015a). On the dynamics of topic-based communities in online knowledge-sharing networks. In *Proceedings of the 2nd European Network Intelligence Conference*, ENIC '15. IEEE.

Guimarães, A., Silva, A. P. C., and Almeida, J. M. (2015b). Temporal analysis of inter-community user flows in online knowledge-sharing networks. In *SIGIR 2015 Workshop on Temporal, Social and Spatially-aware Information Access*, TAIA2015.

Harper, F. M., Raban, D., Rafaeli, S., and Konstan, J. A. (2008). Predictors of answer quality in online q&a sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 865--874, New York, NY, USA. ACM.

Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 462--470, New York, NY, USA. ACM.

Li, B., Lyu, M., and King, I. (2012). Communities of yahoo! answers and baidu zhidao: Complementing or competing? In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. ISSN 2161-4393.

Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J., and Tseng, B. L. (2007). Blog community discovery and evolution based on mutual awareness expansion. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, WI '07, pages 48--56, Washington, DC, USA. IEEE Computer Society.

Ma, W. W. (2012). Online knowledge sharing. *Encyclopedia of Cyber Behavior*, pages 394--402.

Matsubara, Y., Sakurai, Y., and Faloutsos, C. (2015). The web as a jungle: Nonlinear dynamical systems for co-evolving online activities. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 721--731, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., and Faloutsos, C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 6--14, New York, NY, USA. ACM.

Mendes Rodrigues, E. and Milic-Frayling, N. (2009). Socializing or knowledge sharing?: Characterizing social intent in community question answering. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1127--1136, New York, NY, USA. ACM.

Myers, S. A. and Leskovec, J. (2012). Clash of the contagions: Cooperation and competition in information diffusion. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, ICDM '12, pages 539--548, Washington, DC, USA. IEEE Computer Society.

Pal, A., Farzan, R., Konstan, J., and Kraut, R. (2011). Early detection of potential experts in question answering communities. In Konstan, J., Conejo, R., Marzo, J.,

and Oliver, N., editors, *User Modeling, Adaption and Personalization*, volume 6787 of *Lecture Notes in Computer Science*, pages 231–242. Springer Berlin Heidelberg.

Palla, G., Barabási, A.-L., and Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446(7136):664--667.

Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2012). Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3). ISSN 1384-5810.

Radinsky, K., Svore, K. M., Dumais, S. T., Shokouhi, M., Teevan, J., Bocharov, A., and Horvitz, E. (2013). Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM Trans. Inf. Syst.*, 31(3):16:1--16:37. ISSN 1046-8188.

Ravi, S., Pang, B., Rastogi, V., and Kumar, R. (2014). Great question! question quality in community q&a. In *Proceedings of the 8th International Conference AAAI Conference on Weblogs and Social Media*, pages 426--435. AAAI.

Ribeiro, B. (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, pages 653--664. International World Wide Web Conferences Steering Committee.

Ribeiro, B. and Faloutsos, C. (2015). Modeling website popularity competition in the attention-activity marketplace. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 389--398, New York, NY, USA. ACM.

Schoenebeck, G. (2013). Potential networks, contagious communities, and understanding social network structure. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1123--1132, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Solomon, J. and Wash, R. (2014). Critical mass of what? exploring community growth in wikiprojects. In *Proceedings of the 8th International Conference AAAI Conference on Weblogs and Social Media*, pages 476--484. AAAI.

Tang, L., Liu, H., Zhang, J., and Nazeri, Z. (2008). Community evolution in dynamic multi-mode networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, pages 677--685, New York, NY, USA. ACM.

Wang, G., Gill, K., Mohanlal, M., Zheng, H., and Zhao, B. Y. (2013). Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1341--1352. International World Wide Web Conferences Steering Committee.

Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

Zhang, Z., Li, Q., Zeng, D., and Gao, H. (2014). Extracting evolutionary communities in community question answering. *Journal of the Association for Information Science and Technology*.

Zheleva, E., Sharara, H., and Getoor, L. (2009). Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1007--1016, New York, NY, USA. ACM.

Zhu, H., Kraut, R. E., and Kittur, A. (2014). The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 281--290, New York, NY, USA. ACM.