Summer 2014

Online Naturalization: Evolving Roles in Online Knowledge Production Communities

Jeremy David Foote
Purdue University

Graduate School ETD Form 9
(Revised 01/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Jeremy Foote

Entitled
ONLINE NATURALIZATION:
EVOLVING ROLES IN ONLINE KNOWLEDGE PRODUCTION COMMUNITIES

For the degree of Master of Science

Is approved by the final examining committee:

Seungyoon Lee

Lorraine Kisselburgh

Sorin Matei

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Seungyoon Lee

Approved by Major Professor(s): _____

_____

Approved by: Marifran Mattson 07/03/2014

Head of the Department Graduate Program Date

ONLINE NATURALIZATION:

EVOLVING ROLES IN ONLINE KNOWLEDGE PRODUCTION COMMUNITIES

A Thesis

Submitted to the Faculty

of

Purdue University

by

Jeremy D. Foote

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2014

Purdue University

West Lafayette, Indiana

ACKNOWLEDGMENTS

I'm very grateful for the assistance of many people in completing this project. Dr. Seungyoon Lee has been helpful in both inspiring the ideas for this project, and in helping me to think through the implementation and implications. In addition, her suggestions and help during the actual writing process have been wonderfully helpful. Most of what's good about this project (hopefully most of it) is attributable to her aid and direction.

My other committee members, Drs. Lorraine Kisselburgh and Sorin Matei, have also been wonderful. Taking classes from them has inspired much of my thinking, and working with them on multiple projects has provided a great foundation for understanding how research is supposed to be done. Being invloved in the KredibleNet project, led by Dr. Matei, has provided a wonderful foundation for much of this work, and interacting with some of the members of that community has been a highlight of my time here.

I'm thankful to Dallan Quass for making the WeRelate data available to me, and providing support in making sense of it. I'm also thankful for the members of the peer production communities around open source software. I used many great open source tools for preparing, storing, analyzing, and sharing my data and findings, including Ubuntu, Python, PostgreSQL, R, LaTeX, and the R packages RSiena, Hmisc, igraph, statnet, pam, and stargazer.

Above all, I am incredibly thankful to my wife, Kedra Foote, who has been a tireless support, encouraging my studies despite the sacrifices she has made to follow me (2 children in tow) around the country.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Foote, Jeremy D. M.S., Purdue University, August 2014. Online Naturalization: Evolving Roles in Online Knowledge Production Communities. Major Professor: Seungyoon Lee.

Web-based peer production communities, like Wikipedia and open source software, have created digital artifacts of growing cultural, financial, and technological importance. Understanding how and why people choose to join these communities, and why they eventually leave them, is therefore an important topic.

We take all of the edit data from six years of activity on the online genealogy wiki WeRelate, and create monthly snapshots of behavior and interaction networks for all 9,570 users who edited the site. We use machine learning to cluster these behavioral snapshots into four "behavioral roles". We identify one of these roles as being indicative of a community of practice, and we investigate how users move from role to role. As in many other online, peer production projects, the vast majority of users are only active for a short time, and contribute very little while a small number of users contribute a great deal.

Figuring out how to recruit and encourage these users is very important to the success of peer production projects. We use visualizations, regression analysis, and stochastic actor-oriented modeling of four different types of interaction networks to study whether these very active users represent a community of practice that new users can learn from and join. We also study how people leave the community, and whether there are signals that someone is starting to disengage.

We do not find much evidence that these users go through a period of legitimate peripheral participation or acculturation. Rather, those who will become core members show behavior that is similar to long-term core members from their first few

months on the site. We find that these core members show a clear trend of disengaging from the community over a few months before leaving completely, indicating a period where intervention may be effective. We also find a potentially effective intervention, as those who are actively interacting with others who are core members are less likely to disengage.

Our findings provide implications for understanding how online communities function, how interaction networks influence user activity, and how those who are members of these communities might make them more effective. The study also provides a new methodological framework for studying the influence of communicative interactions in online communities.

1. INTRODUCTION

By any measure, the Internet has been one of the most quickly-adopted and pervasive technologies in history. While it is much too early to try to speak about the long-term impact of the Internet, there have already been a number of surprising developments. One of the most surprising has been the scope of what Yochai Benkler calls

> a new modality of organizing production: radically decentralized, collaborative, and nonproprietary; based on sharing resources and outputs among widely distributed, loosely connected individuals who cooperate with each other without relying on either market signals or managerial commands. (Benkler, 2006, p. 60)

Benkler (2006) calls this mode of production "Commons-based Peer Production". While Wikipedia and Linux are perhaps the poster children for commons-based peer production, a number of other projects like Reddit, torrents, and SETI@home are also enabled by peer-production principles. Basically, this includes any project where many participants voluntarily work together to produce a shared digital output. These projects have produced artifacts with far-reaching technological, financial, and cultural effects.

The impact of peer production projects has led researchers to examine how these projects are organized, and why people choose to volunteer their time to work on them. The research on these communities has focused on two primary questions. First, a number of studies have examined what motivates people who are active in these communities (Hertel, Niedner, & Herrmann, 2003; Lakhani & Wolf, 2005; Nov, 2007; Schroer & Hertel, 2009). Second, research has focused on the different roles that people take in these communities, and how people move from one role to another

(Chan, Hayes, & Daly, 2010; Fisher, Smith, & Welser, n.d.; Gleave, Welser, Lento, & Smith, 2009; Welser, Gleave, Fisher, & Smith, 2007; Welser et al., 2011).

We focus on trying to explain two related traits that are common across peer production communities. First, much of the work is done by a small group of core users (Kittur, Chi, Pendleton, Suh, & Mytkowicz, 2007). We see this group of users as a "community of practice", as they share a goal and an understanding of that goal (Lave & Wenger, 1991). We explore why some users become part of this core group, while most do not. Second, we explore why users stop contributing to online communities. While research has been done on why people leave communities (e.g. Baumer et al., 2013), this has been less thoroughly studied.

In peer production communities, people work together toward a common goal, and much of the research views participants as members as of a social network, either explicitly or implicitly. Our study explicitly uses social network analysis as a lens for studying peer production. We build on previous research by analyzing the co-evolution of patterns of behavior (behavioral roles) and four different networks of social interaction: collaboration, observation, local communication, and community-wide communication. We work from an actor-centric perspective, focusing on the local environment around users in order to predict future behavior. In this way, we are able to closely study the interplay between social behaviors and interaction networks.

Our overarching goals are two-fold. First, we want to explore whether the clustering algorithms of machine learning are an effective way to identify behavioral roles in an online community, particularly in identifying core members of a community of practice. Second, we seek to identify how people move through these roles, and particularly how their networks of interaction affect how they change roles.

A greater understanding of these traits of online communities can provide both theoretical and practical benefits. Studying online communities provides insights that can be applied in other contexts. While large-scale online communities are very different from offline communities in many ways, there are many offline volunteer-

based collaborative groups, such as investor clubs or software user groups where there are similarities in motivation and expectations of members. In these offline communities, data collection is much more difficult to obtain, especially social network data. One of the limitations of traditional social network analysis is that network data must be gathered via intrusive and often labor-intensive methods, such as surveys or interviews. This difficulty means that these studies rarely include more than a few dozen participants (Howison, Wiggins, & Crowston, 2011).

In many online communities, on the other hand, every interaction, between every participant, is recorded automatically and unobtrusively. In some communities, such as Wikipedia and WeRelate, these interactions are publicly available for research. This "digital trace data" provides a scope and granularity of data which is simply not feasible to gather in face-to-face communities, allowing for detailed analyses that can include thousands or even millions of participants. Trace data also avoids the problems associated with self-reported relational measures, such as faulty recollection or incomplete networks (Marsden, 1990). We expect that the broad lessons from examining online networks will apply to all sorts of collaborative communities, both online and offline.

In addition, these data sources allow for empirical and *in situ* testing of theories of social interaction and group formation, such as Social Identity, Self-Categorization, and Situated Learning (Hogg & Terry, 2000; Lave & Wenger, 1991). Most previous research in these areas relies on experiments (e.g. Abrams, Wetherell, Cochrane, Hogg, & Turner, 1990) or observation (e.g. Henning, 1998; Lave & Wenger, 1991). Digital trace data allows for a new way of examining these theories via completely unobtrusive collection of fine-grained data, and *ex post facto* analysis of behavior and interactions. In this paper, we provide a framework for translating some of the ideas of Situated Learning into behavioral and network measures. This work can be extended to other related theories.

The findings from this research can also be applied for those seeking to create collaborative atmospheres, either online or offline. For example, these findings offer

software designers and community managers information on how to identify conditions that lead to increases or decreases in participation. This allows them to improve the software that is used for the collaborative projects, as well as to develop tools to monitor and improve community cohesiveness, identify individual "at-risk" users, and increase community productivity.

In Chapter 2, we review the related literature, focusing on online peer production, community roles and evolution, and social network analysis. In Chapter 3, we discuss the context of this paper, and the methods we use to explore the proposed questions. In Chapter 4 we will provide our results. Chapter 5 is a discussion of our results, as well as some caveats about our findings and some ideas for future work. In Chapter 6 we will provide a summary of the project.

2. BACKGROUND AND LITERATURE REVIEW

2.1 Peer Production Communities

Peer production projects have been touted as very inclusive and participatory forms of culture production (Benkler, 2006). They often promise, either explicitly or implicitly, an egalitarian model of participation. The English Wikipedia, for example, is subtitled "the free encyclopedia that anyone can edit", implying that the content on Wikipedia is the result of "anyone" editing content that they are interested in. While networked peer production certainly provides more options than traditional media for people to participate in its production, researchers have found that the distribution of participants' contribution levels in these projects is anything but equal. That is, while anyone (with an Internet connection and a moderately advanced level of technical skills) *can* edit Wikipedia, not just anyone *does* edit Wikipedia. Rather, the contributions follow the Pareto principle, with a power-law-like distribution. The majority of contributors provide very few contributions, while a small group does much of the work, especially the technical and norm-based work (Kittur et al., 2007; Swartz, 2006).

Much research has been done in trying to figure out who does the work on these projects, and why they do it. In every case, there is a group of core contributors who provide the backbone for the project. Research has been done into what motivates contributors (Nov, 2007; Schroer & Hertel, 2009), and how people move from passive reading to active participation (Preece & Shneiderman, 2009). In addition to these user-centric examinations of participants, there is research into what sorts of effects this distribution of participation has on the loci of power, as well as the way that the communities as as whole develop and mature. Although it is natural to see this skewed participation distribution as problematic, some recent work has suggested

that this inequality may actually be an indication of structure and organization, and that equal participation might not be able to sustain or support complex projects like these (Britt, 2014; Matei et al., in press). However, this structure also leads to a model which gives early core members of a community a large amount of influence in selecting the norms and goals of the project, which are not easily changed as the project matures (Shaw & Hill, 2014).

Given the importance of these core members, figuring out how to recruit and maintain more core contributors is of great importance for all peer production projects. Our research seeks to both illuminate why people move into this "inner circle" of very active users as well as to provide some potential ways of encouraging more users to contribute high volumes of work.

2.1.1 Roles in Online Communities

In retrospect, the wide distribution of participation rates should not be surprising. Some of the earliest research into online communities found that people interact in different ways, playing different roles. These roles don't only represent different levels of activity, but completely different uses of the community space, and perhaps different understandings of its purpose. For example, John Seabrook participated in and wrote about The WELL, an early online community. He described a number of roles that he observed, such as "lurkers" and "flamers" (Seabrook, 1998).

However, these roles are rarely clearly delineated. In general, contributors do not hold the same sorts of formal positions that one might find in offline voluntary organizations (e.g, President, Treasurer, etc.). Researchers are starting to build a new understanding of online communities, recognizing that positions of power and influence exist, but are obtained and enacted differently than in offline contexts. Our paper has been inspired by much of this work, which uses digital trace data and interaction networks as a means of understanding how these communities function (Bertino & Matei, in press; Matei, in press; Smith, Rainie, Shneiderman, & Himel-

boim, 2014). These researchers have found that peer production communities function as something of an "adhocracy", where contributors gain power and influence via their activity in the project, rather than through top-down assignation (Matei et al., 2014).

Authors examining peer production communities and discussion boards have identified roles such as "readers", "contributors", "collaborators", and "leaders" (Preece & Shneiderman, 2009), "Popular Initiators", "Taciturns", "Popular Participants", and "Elitists" (Chan et al., 2010), and "Substantive Experts", "Technical Editors", "Social Networkers", and "Vandal Fighters" (Welser et al., 2011). These roles are sensitive to context, and while some roles might exist across many kinds of communities (e.g., "readers"), others will be different in different communities.

The different roles that users take do not simply indicate an effective way of distributing work, but give evidence of a complex social structure. In role theory, people are seen to adopt behaviors that meet the expectations of others (Merton, 1957; Mead, 1934; Goffman, 1959). Where we find social roles, we find communities with power structures and diverse kinds of relationships. Merton, for example, argues that each social role is defined by a "complement of role-relationships", which are expectations about how to interact with others based on their roles (Merton, 1957, p. 110). Thus, social roles include both a behavior component and an interaction component. For example, those who have the role of "new parent" would probably spend more money on diapers and bottles (behavior), sleep less (behavior), and spend less time with friends (interaction) than those with the role of "newlywed" or "single person". Roles are culturally contingent, and can be considered as patterns of behavior and relations with others (Gleave et al., 2009). Although roles are persistent, they are not permanent, and can change over time.

2.1.2 Situated Learning in Online Communities

We use Lave and Wenger's (1991) theory of Situated Learning as a framework for understanding how and why people move through different roles over time. *Situated*

Learning was the title of Lave and Wenger's seminal book, in which they introduce two key concepts: communities of practice and legitimate peripheral participation. Communities of practice are groups of people who "participat[e] in an activity system about which participants share understandings concerning what they are doing and what that means in their lives and for their communities" (Lave & Wenger, 1991, p. 11). Communities of practice are groups of practitioners, and provide a means of transmitting applied knowledge, through relationships. They are generally fluid in both their membership boundaries and scope, and individuals may be a part of many different communities of practice (Lave & Wenger, 1991).

The shared understandings and purpose which define a community of practice must be learned, through observation and "legitimate peripheral participation" (Lave & Wenger, 1991). Legitimate peripheral participation involves learning by doing. Initiates participate in activities that contribute to the overall goals of the community, but are not particularly vital to the community. As they perform these tasks, they not only learn necessary skills, but also the norms and beliefs necessary for core participation in the community. Lave and Wenger (1991) give the example of apprentice tailors, who begin with simple tasks like sewing buttons or maintaining sewing machines, and move into more complex and more important tasks like cutting the cloth.

Although the theory was initially developed before the Internet became widespread, researchers have applied it to online communities, including peer production communities. Preece and Shneiderman (2009), for example, provide a framework for how people move into leadership roles in online contexts, with examples from a number of communities.

Bryant, Forte, and Bruckman (2005) explicitly use Situated Learning as a framework for understanding how users become active on Wikipedia. Interestingly, they claim that many participants become involved in tasks that could be considered legitimate peripheral participation before recognizing that there is any sort of organization that is behind the project. Because reading or making simple edits can be done

without communication or collaboration, new users assume that the entire project is simply the aggregation of small contributions made by unconnected strangers. The active Wikipedians interviewed describe how even after making a number of edits, they didn't realize that there was a "community" at all. Thus, the first step toward joining the community is recognizing its existence. As predicted by Situated Learning, these Wikipedians report moving into community leadership roles only after a period of peripheral activities, such as making simple edits to articles they were interested in. Over time, Bryant et al. report that users moved from working independently on subjects of interest in Wikipedia to doing more community-centric tasks, such as organizing content or discussing policies. Just like the tailor apprentices, with exposure to the community, they learned the skills, norms, and beliefs necessary to achieve core participation.

We apply this framework of Situated Learning to predict how and why members move through behavioral roles over time. Specifically, we expect people to perform peripheral activities as they begin their membership in a community, but over time those who remain in the community will increase their overall activity in the community, their interaction with other community members, their participation in tasks that require specialized skills or knowledge, and their participation in community-level tasks such as maintenance or policy discussions.

2.2 Social Network Analysis

Situated Learning is an inherently social way of learning, and social network analysis provides a natural and powerful way of exploring this theory. Social network analysis includes "theories, models, and applications that are expressed in terms of relational concepts or processes" (Wasserman & Faust, 1994, p. 4). Social network analysis recognizes that many contexts (such as communities of practice) cannot be understood without examining both the behavior of participants as well as the relationships between them.

Despite what observers or casual participants might think, most of the work that is done in peer production communities is collaborative work, and there are meaningful relationships between participants. This collaboration can take a number of forms, and we use social network analysis in order to analyze the co-produced nature of behavioral roles and interaction networks.

Indeed, peer production researchers, and particularly those studying Wikipedia, have often used social networks analysis in their research. For example, these tools have been used as a way to identify polarizing articles and editors (Brandes, Kenis, Lerner, & van Raaij, 2009), to look at how users choose collaboration partners (Keegan, Gergle, & Contractor, 2012), to try to predict the nature of the relationship between users (Leskovec, Huttenlocher, & Kleinberg, 2010), to study how users collaborate to produce new information (Keegan, Gergle, & Contractor, 2013), to detect patterns of collaboration (Iba, Nemoto, Peters, & Gloor, 2010; Laniado & Tasso, 2011), and to study system-level properties of these communities (Britt, 2014; Crowston & Howison, 2005; Matei et al., 2014; Voss, 2005). Social network analysis has also been used as a tool in identifying social roles (e.g. Chan et al., 2010; Gleave et al., 2009; Welser et al., 2007, 2011). We build on this literature by developing new methods for creating interaction networks, applying social network analysis to social role evolution, and applying methods used primarily in Wikipedia to a new context.

2.3 Hypotheses

We bring together the concepts of roles, Situated Learning, and social network analysis in order to study how people become part of an online genealogy community, and why they eventually leave. We begin with some foundational hypotheses about behavioral roles and Situated Learning.

2.3.1 Behavioral Roles

In a paper that provides a foundation for our work, Welser et al. (2011) examine behavioral patterns and network "structural signatures" as indicators of roles. They follow the recommendation of Gleave et al. (2009) by starting with socially important roles identified qualitatively. "Substantive experts" are experts about topics that they edit, and they are the people who actually add most of the content to Wikipedia. "Technical editors" correct small errors across the site and organize and standardize the project. "Vandal fighters" use a number of tools to stop or correct the work of vandals on the site, and "Social networkers" focus on interaction with other users, and on building community. Welser et al. (2011) examine the editing behavioral patterns and egocentric talk networks of a small number of hand-picked users, representative for each role. They use the observed behavioral patterns for each role to create rules for classifying other users into these roles. For example, they classify technical editors as those with ">60% of total edits to content pages, >45% [of non-content edits] in Wiki and Infrastructure combined, and <25% [of non-content edits] in content talk" (Welser et al., 2011, p. 127). Substantive editors had more content and especially content talk edits, while social networkers had lots of user and user talk edits and vandal fighters had lots of content and user talk edits.

They then use these classification rules to identify users who fit into each role, in two different groups – a set of dedicated editors, and a set of new editors, with the goal of comparing role prevalence in the two groups. They find that the distribution of roles is similar in both groups, with substantive experts and technical editors being the most common roles.

We take a different approach to finding roles. While other research generally identifies roles based on observation and intuition, we do not start with qualitatively identified roles. We propose that machine-learning based clustering algorithms may be applied to behavioral data, which allow for automated discovery and classification of user roles. We take aggregated behavioral data for users, and cluster them into

these behavior-based roles, and then use qualitative analysis to interpret the results. We then examine the network signatures for each of the identified behavioral roles, for each of four different interaction networks, to try to identify distinctive patterns of interaction within these networks (cf. Stark & Vedres, 2006).

Machine learning refers to computer programs that improve as they get more data. While there are a number of different problems that machine learning can be applied to, we use it for clustering behavioral patterns. According to Xu and Wunsch (2005), "The goal of clustering is to separate a finite unlabeled data set into a finite and discrete set of 'natural,' hidden data structures" (p. 645). For this paper, the "hidden data structures" that we are seeking to uncover are behavioral roles, and we use clustering algorithms to group behavioral patterns into these roles. By using clustering, we may be able to identify patterns of behavior that would be difficult or impossible to discern simply by observing a community.

Our primary goal is to identify clusters related to roles in a community of practice. Situated Learning theory suggests that members of a community start out on the periphery, but over time they learn the skills, norms, and beliefs necessary to become core participants. If, as Lave and Wenger (1991) suggest, members of a community of practice share beliefs about the meaning and purpose of the community, and their roles in it, then it is reasonable to believe that they should display similar behavioral patterns, and therefore would be clustered together by a clustering algorithm.

Our hope is that we will be able to identify automatically the sorts of roles that Welser et al. (2011) are able to identify through observation of the community, as well as identifying other roles which might not be as visible to outside observers. In addition, this sort of analysis allows for the automated identification of which behaviors and network structures are salient for differentiating roles.

That being said, this technique for identifying behavioral roles is novel, and we may have difficulty in identifying some of the subtle differences between roles. Whether or not we are able to identify all of the anticipated roles, we seek to at least cluster users into a few broad categories, where one or more of the categories represent peripheral

participation, and one or more represent core participation in the site as a community of practice.

We anticipate that the behavior that some users engage in will fit with what we would expect of core members of a community of practice: they will be more active in the community, and in particular, they will participate much more in community maintenance and community discussion tasks.

Formally, we wonder:

RQ1. Will some behavioral roles fit the profile of core membership in one or more communities of practice (i.e., editing more often in the community, doing more complex tasks, and communicating more with other users)?

2.3.2 Roles and Networks

Once we have identified roles which represent peripheral community membership and core membership, we examine how members with these various roles interact with each other. Social network analysis allows us to analyze how behaviors are related to interaction networks. While most studies of peer production communities focus on co-editing relationships (e.g. Britt, 2014; Brandes et al., 2009), we create four different types of interaction networks, based on Situated Learning and the findings of Bryant et al. (2005).

Situated Learning describes how observing the work of core participants can be important for learning the norms and skills needed to become part of a community of practice, so we include an observation network. We also include a collaboration network, intended to identify times when users directly work together to create knowledge, and a local communication network, which identifies opportunities for direct transfer of knowledge or norms. Finally, based on the observations by Bryant et al. (2005) that core members of a community interact more at a community level, we create a network that specifically tracks community-level communication. Creating multiple types of networks (or "multiplex" networks) can provide valuable insights

into how communities work, which might be lost when just looking at one network (Lee & Monge, 2011). Details about how these networks are created is given in the Methods chapter.

As mentioned, one of our goals in assigning roles based only on behavior, and not on network position, is to examine how social position differs by these behavioral roles. We assume that different behavioral roles will also have different "network signatures", for each of the interaction networks. Because our method of creating both behavioral roles and interaction networks is novel, our general goal is exploratory, and exactly how network statistics will differ by roles is an open question.

RQ2. Will behavioral roles have unique network signatures for each type of interaction network?

While the broad question of how network signatures will differ based on behavior roles is difficult to predict (most pointedly because we don't know what sorts of roles we will find!), assuming that we identify roles associated with core membership in communities of practice, the literature can guide us to make a few predictions about their network signatures. First, they should be involved in more interactions than others. Not only should they interact with more people, but they are likely to interact with other core members, about central issues, more than they interact with peripheral members (Bryant et al., 2005).

This interaction pattern predicts three social network results. First, those whose editing behavior fits the role of core membership in a community of practice will be central in interaction networks. Centrality is a measure designed to assess how well-connected a given node is to the other nodes in a network. There are a number of different measures of centrality, and for this paper we use eigenvector centrality. This method differs from other centrality measures by using factor analysis to give less weight to a node's local centrality and more to its overall position in the network (Hanneman & Riddle, 2005). For example, this measure would give a higher score than other measures to nodes that are connected to a few central nodes, even if they

are not closely connected to any other nodes. Core members should be in the center of the network, and should therefore have high eigenvector centrality scores.

H1. Members of roles representing core members in the community of practice will have higher eigenvector centrality scores than those in peripheral roles, for all interaction networks.

In addition to this broad prediction about all interaction networks, if the core members follow the same sort of path described by Bryant et al. (2005), then we should expect them to participate much more in community governance and decision-making than other users. If they are participating much more in this role than others, then we would expect them to be most central in the network representing community talk.

H2. The difference in eigenvector centrality between core members and others will be greatest in the community talk network.

The idea that the core members will interact more with other core members can also be tested through examining these interaction networks. First, we would predict that there are cliques – that is, that core members are in groups where everyone knows everyone else. This can be tested by looking at how many of their interaction partners interact with each other – that is, how "clustered" these users are in the interaction networks.

H3a. Core members will be in more tightly clustered groups than others, for all interaction networks.

Relatedly, we predict that core members are not only more likely to be in cohesive groups, but that those groups will be composed of other core members. There are a number of algorithms for creating "communities" of nodes in a social network, and we look at whether these network-based communities of users are composed of those who are in the same behavioral role. Core members should interact with each other often, and should therefore often be put in the same network community as each other, while peripheral members will be spread across the various network-based communities.

H3b. Core members will be more likely than other roles to be in the same network-based communities as each other.

2.3.3 Longitudinal Analysis

For our initial hypotheses, we examined the aggregated roles, and aggregated networks of users. This can give a useful overall picture of a network, but in order to understand how behavior and interaction networks co-create each other, we must use longitudinal network and behavioral data. We first make predictions about the evolution of behavioral roles independently, and then about how they interact with interaction networks.

Role Evolution

For new members of an online peer production community, there are very few barriers to leaving the community, and the distribution of both edits and length of tenure bear this out, with the vast majority of participants contributing little and quitting early (Kittur et al., 2007). This provides evidence that the early experiences that people have in these communities have a big impact on their future participation, so we look at the early behavior of all of the participants of the site, to see if the roles that one takes early on can help to predict whether they become a core member of the community in the future.

The literature predicts two ways that those who later become core members differ from those who do not. The first seems to partially contradict the idea of communities of practice. Panciera, Halfaker, and Terveen (2009) found that those who become highly active on Wikipedia display unique behavioral patterns from the very beginning of their participation. For example, they have a much higher mean number of edits, even in their first few days of editing, and they are much more likely to edit every day. Interestingly, participation in norm-based or skilled tasks (such as edits in the "Wikipedia" namespace or edit comments that reference Wikipedia policy) starts low

for both types of users, and grows quickly for the highly active users (Panciera et al., 2009). However, this finding can be reconciled with the idea of communities of practice, if we assume that while those who eventually become core members start differently than others, they will still follow a detectable pattern of learning.

Our second prediction tests this idea. We anticipate that as users first start using the site, they will initially be in peripheral roles while they learn the skills and norms necessary for core participation. As mentioned, Preece and Shneiderman (2009) identify a progression of roles through which some members of a community pass, on the way to becoming "leaders", which is similar to our operationalization of core members in a community of practice. From this perspective, many of the roles in a community can act as a stepping stone to other, more core roles.

We hope to find evidence of these two aspects, by examining the first few months that users are on the site. With this goal, we take the behavioral roles which we identified based on behavioral measures, and then identify the early temporal role paths that users pass through on their way to becoming active, core members of a community.

H4a. Users who become core participants will participate in different behavioral roles than others from the very beginning of their time on the site.

H4b. Users who eventually move into core participant roles will begin their tenure on the site in peripheral roles.

While we have focused so far on what conditions encourage peripheral participants to become core users, it is also important to understand why and how people leave the community. There are a number of different reasons that people leave online communities, such as the community diverging from them (Danescu-Niculescu-Mizil, West, Jurafsky, Leskovec, & Potts, 2013), exhaustion, discontent toward other participants, or privacy concerns (Reagle, 2012). We hope to tease out role evolutions that identify some of these reasons as they are being enacted. If these pathways are identifiable, then this gives a way to identify users who are in danger of leaving the community, and who may benefit from some sort of intervention. We believe that that just as

the sequence of roles that leads to becoming a core member will follow a predictable pathway, users who drop out of the community will be similarly identifiable.

H5. The recent behavioral roles that users have been in will predict whether they will stop participating in the community.

Social Influence on Behavior

Having examined the role pathways simply by looking at the behaviors that users exhibit over time, we next analyze the relationship between behaviors and interaction networks. Our goal is to try to figure out what sorts of interactions with others lead users to adopt various behavioral roles. As users engage in peripheral activities on peer production sites, the amount of contact that they have with core members is partially due to luck. For example, a new user may happen to be interested in the same topic as a core user, and their paths will cross. This "old-timer" can influence the new user to become more active in a number of ways, such as: sending encouraging messages, helping to teach technical skills, introducing the new user to other community members, or modeling community norms.

This passing of behaviors via a network tie is called "social influence" or "social contagion". Whether someone adopts a new product (Aral, Muchnik, & Sundararajan, 2009), what sorts of health decisions they make (Christakis & Fowler, 2007), and even what their emotional state is (Fowler, Christakis, Steptoe, & Roux, 2009) can all be influenced by the social networks one is a part of.

We use stochastic actor-oriented models to examine a few different ways that core participation might be "contagious", by examining whether core participant behavior spreads through each of the interaction networks that we have identified. First, we focus on observation. Observing the work of core participants can help others to become integrated into the community of practice (Wenger, 1998). We expect that by observing the high-quality, well-researched pages that core members have created,

new or peripheral users will be encouraged to learn the skills necessary to do the same.

H6a. Users who observe the work of core participants will be more likely to move into core participant roles.

Second, we anticipate that collaboration will be a powerful way of becoming part of the community. In contrast with observation, collaboration includes both observation and repeated interaction with experienced users, as both members work together on an artifact. Collaborations should be opportunities for correction and instruction, where new users gain skills and learn norms necessary for successful future collaborations. It may be through collaboration that users are first made aware of the community structure of the site and the negotiated nature of the artifacts (like the Wikipedians in Bryant et al., 2005).

H6b. Users who collaborate with core participants will be more likely to move into core participant roles.

Third, norms, advice, and encouragement can be communicated to peripheral users via the communication channels offered on the site. For example, there are a number of users who look for errors or problems, and communicate with new users to resolve them. The following is a typical message from one of these users:

> Hello, we noticed that you've recently added several individuals to WeRelate with numbers in the name field, such as Thomas McClelland. We'd ask that you please do not add any additional numbers to the name field of persons on WeRelate and also would appreciate it if you could remove the existing ones, as numbers and/or special characters in the name field may interfere of cause errors in the search and/or merge functions at WeRelate, per the WeRelate Naming Guidelines as listed on the Help:Person Pages Tutorial.
>
> Thanks for your assistance and best regards,

Jim Volunteer Administrator, WeRelate

Choi, Alexander, Kraut, and Levine (2010) found that messages to new users on Wikipedia increased the likelihood that they would continue to contribute. We anticipate that a similar effect would occur in this context, and that those who receive and respond to these sorts of messages will be more likely to develop the knowledge and relationships necessary for core participation.

In addition to these personal messages, the site provides spaces for more general communication, where community-based discussions can be held. We anticipate that participating in these community-wide discussions will also lead to increased learning and acculturation, and we look for the influence of both types of communication.

H6c. Users who communicate with core participants will be more likely to move into core participant roles.

In addition to directly picking up behavior from those one interacts with, simply being well-connected to the community may influence people to change their behavior to be more active in the community. Eigenvector centrality is an especially useful measure of centrality in this context because it should give higher scores to those on the periphery (i.e., those not connected to many people) who have contact with core participants (i.e., those central in the network) than those connected only to other peripheral members. Those with high eigenvector centrality scores have greater access to the expertise and cultural knowledge of the community – resources which are important for becoming a part of a community of practice.

H7. Users who have high eigenvector centrality in interaction networks while in peripheral roles will be more likely to advance to core participation roles.

Tie decay

Finally, we introduce network interactions into our study of those who quit the community. The way that users are connected to others in a network can affect not only whether they move into core participation, but whether they continue to

participate at all. We begin by studying local decay; that is, the reasons that a relationship between two users does or does not persist over time. Understanding which users are likely to continue interacting with each other can help us to understand which relationships are most important to the users (Burt, 2002). It can also give a deeper understanding of the lived nature of the behavioral roles that we identify.

When looking at a community of bankers, Burt (2000) found a number of different factors that affect the likelihood that a relationship will decay, including length of relationship, homophily, and connectedness. We focus on studying the effect of homophily. In many types of networks, across many contexts, similarity, or homophily, has been found to predict ties between two people or organizations (for a summary of research, see Monge & Contractor, 2003). This is similar to the idea of assortativity as explored by Britt (2011, 2014), except that we use behavioral roles instead of network position as our measure of similarity.

From this perspective, users with different behavioral roles would be less likely to maintain ties. This could be because they simply generally work in different areas of the site, or because they are dissimilar in other regards, which leads to both dissimilar behavioral roles and a lower likelihood of interacting. We predict that users who share similar behavioral patterns (a.k.a. roles) will be more likely to continue to observe, collaborate, and communicate with each other, due to homophily.

H8a. Ties between members with the same role will be less likely to decay.

While we suspect that homophily will hold for most behavioral roles, in a partially competing hypothesis, we suspect that we may find that certain types of roles are complementary, providing symbiotic benefits to each other. Aldrich and Ruef (2006) define symbiotic relations as "two populations [that] exist in different niches and benefit from the presence of the other" (p. 247). For example, some users in an online knowledge production community might focus on the technical aspects of the site, while others have substantive knowledge about the topics of interest. These users may contact each other for help with their respective areas of expertise, and their interactions will be both mutually beneficial and difficult to replace. These different

approaches may be captured by different behavioral roles, in which case we would expect some types of heterogeneous ties to be less likely to decay.

H8b. Ties between members with different roles will be less likely to decay.

Community Departure

People who are embedded in a dense relationship network have "social capital" through these relationships, through which they gain social support, as well as access to resources (Coleman, 1988). Not surprisingly, ties between others in a dense community have been found to be more resilient than other ties. For example, relationships between bankers were more likely to continue when both bankers were "embedded" – that is, when they had a number of connections in common (Burt, 2000).

We suspect that the benefits of embedding are more general than just between dyad ties. Or, from a negative perspective, if there is a social cost for cutting ties with someone in a dense network (think of fighting with one person in a group of your close friends), then there will be a much greater cost for completely leaving the group, which would sever all ties. Thus, we predict that individuals who are embedded are less likely to leave the community, independent of how active they have been in the past.

H9a. Users who are embedded in interaction networks will be less likely to quit the community.

Eigenvector centrality can be seen as a separate, but related measure of social capital. Those with high eigenvector centrality have more connections to other well-connected members, and theoretically have more influence over the direction that the community takes, even if their local group is not as tightly embedded. Those with higher eigenvector centrality would also have more to lose by leaving the community, and thus:

H9b. Users who are central in interaction networks will be less likely to quit the community.

3. METHODS

In this chapter, we describe the context of the research, as well as the methods used to collect, pre-process, and analyze the data.

3.1 Context

The community of focus is the WeRelate genealogy community, hosted at `http://www.werelate.org`. Genealogy is the search for information about ancestors and their lives (Yakel, 2004). This search is often undertaken by individuals or small groups of relatives, who gather and store information for personal use, with limited distribution to family members. As noted by Willever-Farr and Forte (2014), genealogy is simultaneously a way of memorializing family members and creating a historical resource. The fact that genealogists are generally related to the people they research leads them to feel ownership of their family trees. Even when genealogists put their information online, they generally expect to maintain control of the information that they upload. For example, the largest genealogy company, Ancestry.com, provides a space for users to create online family trees. They provide tools for sharing, searching, and collaborating on others' trees, but in the end, each tree is owned by a specific individual, who has the ultimate control over the content.

This configuration allows users of these sites to use them as way to memorialize their ancestors, with full control of how that memorialization is enacted (Willever-Farr & Forte, 2014). This approach has some serious limitations, however. First, there is a huge duplication of effort. A search on Ancestry.com for family trees which include my ancestor, Nathaniel Foote, turned up over 14,000 nearly identical results (see Figure 3.1). A related problem is that poorly researched or incorrect information can spread quickly, and can become accepted simply because it has been accepted by

others. These problems both stem from the fact that sites like Ancestry.com allow users to copy portions of others' family trees directly to their own tree. Ironically, by providing this powerful tool for collaboration, the overall quality of the research has decreased. In fact, many of the most capable genealogists often make their trees private, since the general quality of trees is so low (Willever-Farr & Forte, 2014).

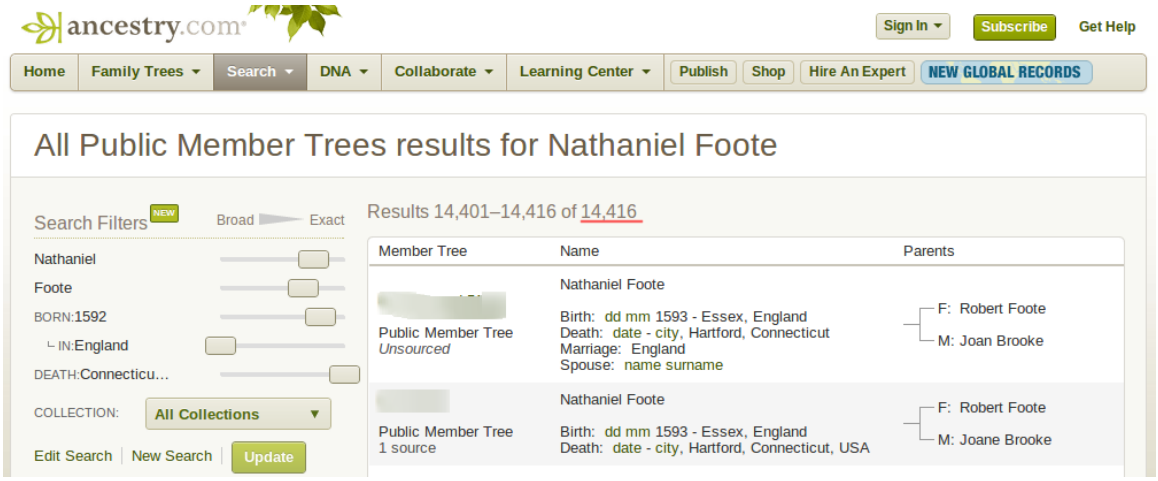

Fig. 3.1: Large number of public trees on Ancestry.com that contain a single individual – Nathaniel Foote

WeRelate provides a very different model of genealogical collaboration, structured after online peer production communities. Instead of working individually, WeRelate members work to build a collaborative online genealogical tree. The site is built on MediaWiki software, the same software used to run Wikipedia (among other sites). Instead of writing pages based on topics, pages are created for deceased people, with relationship links between pages. Like Wikipedia, the site has an open license, allowing for redistribution of content, and making it very easy to study the interactions between users.

Genealogy is actually a common online activity (Yakel, 2004), but is underrepresented in the literature, especially the network literature. This may be because genealogists are generally much older and less tech-savvy than typical Internet users (and the researchers who study Internet users). Studying a genealogy community

provides a useful opportunity for testing the generalizability of theories which were developed studying open-source software communities or Wikipedians to a peer production community with very different demographic characteristics. In addition, the mode of working on WeRelate is quite different. While Wikipedians are generally guided by interest in the topics they edit, genealogists generally work to research the people they are related to, and particularly their ancestors. This has the potential to cause some interesting social dynamics. If you don't get along with another Wikipedia user, then it's usually simple to find another topic of interest to edit. Your ancestors don't change, though – you cannot simply find another grandfather to study if you disagree with your cousin's conclusions.

3.2 Data Collection

On WeRelate, every change to the site is recorded, together with the ID of the user who made the change. Our data comes from an XML dump of all of the edits made on WeRelate from the beginning of the site in January of 2006 to February of 2013. The content of edits were removed, and the remaining data was put into a PostgreSQL database. This data includes the timestamp of the edit, the page that was edited, and the user who edited the page.

We first identified a number of edits that were automated, and removed them from the analysis. These edits were of three main types: First, a bot account was used to automatically create supporting pages from another database. These consisted of "Place" pages and "Source" pages. WeRelate pages about a person generally include events (birth, marriage, death, etc.), both to uniquely identify a person, and to provide a story about their life. These events can be tied to places and/or sources, and can be linked to "Place" or "Source" pages, which describe the place or source, respectively.

The second type of automated edit is created when a GEDCOM file is uploaded. GEDCOM is the de facto standard for transferring genealogy data between programs, and WeRelate allows users to import files that they have created in other software,

via uploading a GEDCOM. When a GEDCOM is uploaded to WeRelate, the system automatically creates pages for each of the people in the file. Because these pages do not represent manual interaction with the site or with other users, we also ignore these edits.

The third type of ignored edits are those that automatically propagate to other pages. For example, "Family" pages include a list of the mother, father, and children in a family, together with their birth and death dates and locations. All of this information comes from "Person" pages, and if any of the relevant data changes on a person's page, then the changes are propagated to any family pages that they appear on. For example, if a user changed the birth date of a person, then the corresponding "Family" page for that person's parents and siblings would be automatically updated. We keep the original, manual edit to the birth date, but ignore the automated edits, to avoid double-counting.

Having access to the entire edit history allows us to analyze the community as a whole. Once we have removed the automated edits, we are left with 2,679,625 edits, made by 9,568 different users. In Figures 3.2 and 3.3, we look at the general trajectory of the site. First of all, we see very low activity at the beginning of the site. Looking directly at the types of edits that occur in 2006, we found that most of them were done by automated bots, or by a very small group of users. The actual launch of the site wasn't until 2007, so we begin our analysis of the site at that point (Quass, 2014).

The rest of the graph is fairly stable. This is very different from the graphs of Wikipedia's early activity, for example. Instead of exponential growth, we see about the same number of active users, and the same number of new users, making about the same number of edits, for years, with a moderate increase in both users and edits starting in 2011. This stability of the number of active users provides some evidence that this is a community of practice, with an established group of core participants.

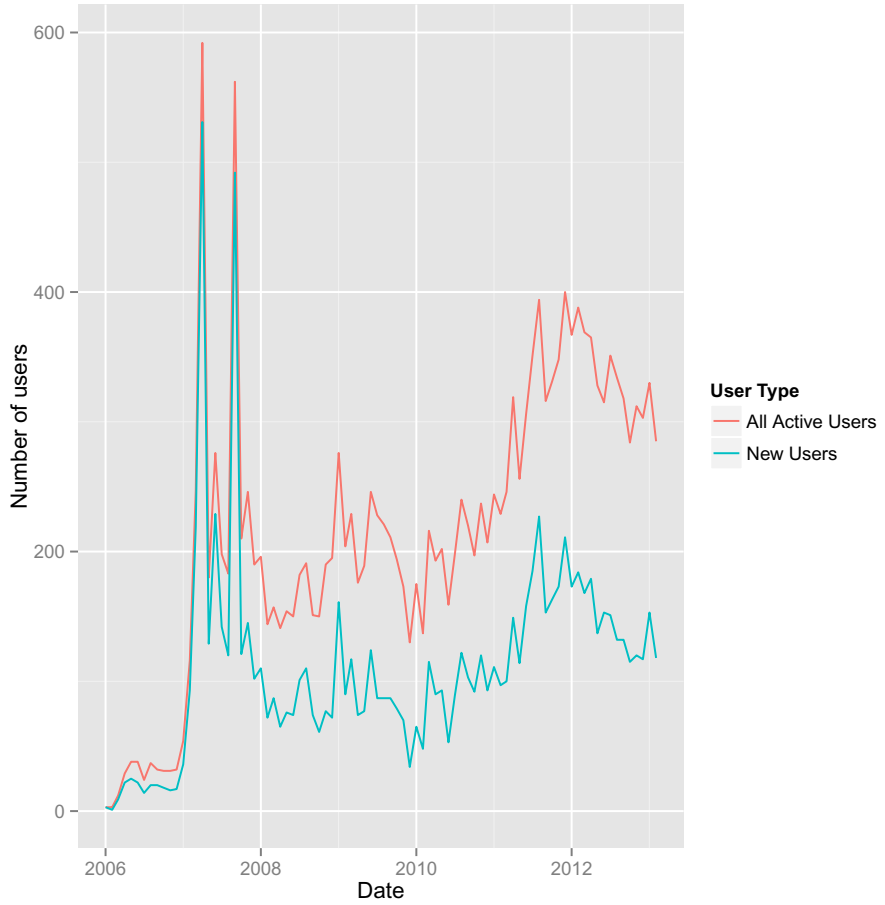

Fig. 3.2: Number of new users and number of active users per month, over full lifetime of WeRelate.

3.2.1 Metadata

After the automated edits were removed, we used the remaining edits to create metadata to measure different aspects of behavior. We extracted the statistics described in Table 3.1 for each user, calculated for each 30-day period, starting 1 January 2007. To simplify the sorts of edits made, we combined a number of different edit types into five higher-level categories: "Simple", "Complex", "Local Talk", "Community Talk", and "Community".

We chose thirty days in an attempt to balance our desire to find stable, meaningful roles and relationships with our desire to detect evidence of learning and change. Since

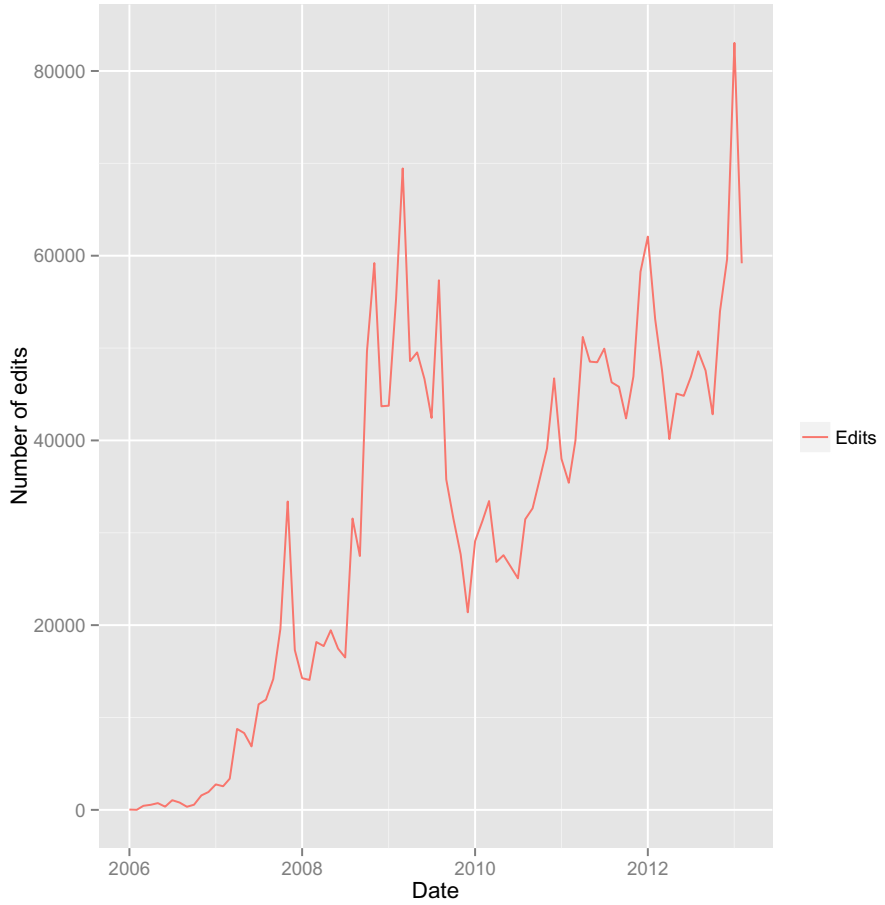

Fig. 3.3: Number of manual edits per month, over full lifetime of WeRelate

many editors only make a few edits per month, having these longer periods allows us to capture this sporadic activity.

These statistics were stored in a table in the database.

3.2.2 Network Creation

The edit data was also used to create four different networks, each capturing a different type of relational behavior between users. This is similar to the approach used by Laniado, Tasso, Volkovich, and Kaltenbrunner (2011), who created multiple net-

Table 3.1: Summary of statistics measured

Time Variant Measures	Composed of
Total number of edits	Manual and automated edits
Manual edits	Manually created edits only
Percentage of "Simple" edits	Edits made in Person and Family namespaces
Percentage of "Complex" edits	Edits in Place, Category, Given Name, Image, MySource, Portal, Repository, Source, Surname, Template, Transcript, and User namespaces
Percentage of "Local Talk" edits	Edits in Family Talk, Person Talk,Given Name Talk, Place Talk, Image Talk, My Source Talk, Repository Talk, Source Talk, Surname Talk, Article Talk, Template Talk, and User Talk namespaces
Percentage of "Community Talk" edits	Edits in Category Talk, Help Talk, Portal Talk, WeRelate Talk namespaces
Percentage of "Community" edits	Edits in Help and WeRelate namespaces
Percentage of edits in all other namespaces	
Time Invariant Measures	
Date of first edit	
Date of most recent edit	

works based on different types of talk pages on Wikipedia, although we operationalize our networks differently.

As with behavioral roles, we created these networks in 30-day windows of time. However, because of the computational complexity in running the stochastic actor-oriented models that we use to examine the interaction between roles and networks, we only create networks for the final eight months on the site, and with only the more active users. We do this by identifying all of the users whose first edit occurred before 1 June 2012, and who made at least 5 edits in two or more 30-day periods from 3 June 2012 to 28 February 2013. This results in a list of 161 users. We then capture the networks that involve only these nodes in the eight 30-day periods starting on 3 June 2012. This process should eliminate only those users who are completely peripheral to the network, and who we would not expect to influence or be influenced by others.

Observation Network

First, we create an observation network. In this directed network, an edge exists from i to j if i has observed j's work. Since most pages on the site are fairly small, we assume that edits made by one user will be observed by the next user to view or edit the page. Unfortunately, data on which users view which pages is not available, so edges are based only on edits. Specifically, for each edit e, an observation is assumed to have occurred between the editor i and the person who made the last non-automated edit j, as long as j is not i. The observation is considered to have occurred at the time of i's edit. Observation can occur within any context on the site, so all types of pages are included. Learning from observation can also occur regardless of the distance in time between edits, so we don't restrict observation based on time. If the person (j) who made the last edit is not in our list of active nodes, then no tie is created. It is worth nothing that this restriction means that we may miss out on the influence of observing those who have left the community, or are no longer active enough to be in our list.

Collaboration Network

We also create a collaboration network. This is an undirected network, with edges between nodes who have collaborated. Our operationalization of a collaboration has two components:

1. Two or more editors have to provide alternating edits. That is, an edit pattern of $\{i,j,j\}$ would not be considered a collaboration, whereas an edit pattern of $\{i,j,i\}$ or $\{i,j,j,i\}$ would each create a single tie between users i and j, while the pattern $\{i,j,k,i\}$ would create ties between i and j, as well as between i and k, but not between j and k.

2. The two edits by i must have occurred within 30 days of each other.

Many other papers focus on co-editing networks, where directed or undirected ties are created for editors who both work on the same page in the same time period (e.g. Britt, 2014; Keegan et al., 2012, 2013; Iba et al., 2010; Laniado & Tasso, 2011). We believe that our measure provides a nearer approximation of what Wood and Gray (1991) call the "interactive process" of collaboration. This term indicates a recursive relationship, with both parties contributing in turn. It is this interactivity that our measure of collaboration tries to capture.

Local Communication Network

Lastly, we identify two types of communication networks, representing the two primary types of communication that occur on WeRelate.

The local communication network is defined by interactions that occur either on a user's Talk page or a content page's Talk page. A sample of recent changes to these sorts of pages showed that they are used primarily for coordinating content creation, but also include advice-seeking, correction, and debate. All of these contribute to the theorized community of practice.

After identifying the namespaces that are primarily local communication about content, we create a network using those namespaces. For talk pages that are on content pages, we define a tie as occurring when i and j both edit the same talk page within 7 days of each other, with the latest edit occurring during the current 30-day window. We hope that this will capture two-way communication, versus simply two people independently writing notes on a page. On User Talk pages, communication is more complicated, and it can happen either by both users writing on the Talk page of one user, or by writing on each other's talk pages. We capture both of these types of communication, by examining both talk pages. A tie exists if i writes on j's user talk page, and then j either also writes on her own talk page, or writes on i's talk page within the next 7 days. This allows us to only capture talk that is two-way; messages which are read but not responded to are ignored.

Community-wide Communication Network

The second type of communication network occurs at the community level. WeRelate has a number of community-wide pages. These pages are used to identify and discuss community norms and standards, ask general questions, and provide Help and FAQ pages to support new members. Defining communication is not as clear on some of these pages. For example, the Watercooler page is designed as a space to discuss myriad community-related topics. Thus, even sequential edits may be on completely different topics, and users interested in one topic may not have even read information in other topics. For these pages, we take advantage of the edit summaries provided by users. When an edit is made to a topic on a page, by default the edit history contains the name of the topic that has been edited, set off by the textual markers "/*" and "*/", e.g., "/* Promoting the Community */". We identify these topics, and treat each topic as a separate page, in the sense that ties are only created between users who edit the same topic. In some cases, the topic does not appear in the edit summary, either because the user edited the entire page instead of just one topic, or

because they removed the auto-generated summary when submitting their edit. In these cases, we err on the side of creating ties, and treat the edit as though it were an edit to all categories on the page.

3.3 Behavioral Roles

We begin our analyses by examining the short-term behaviors of users, to identify different behavioral roles. These measures are taken for each user in 30-day periods, which we call "user-months". We use the short-term behavior information listed previously to create a behavior vector for each user, for each time period. These features include (logged) total edits, (logged) manual edits, and the percentages of simple, complex, local talk, community talk, community, and other edits. We use these behaviors to map the "practice space" of the community, and then group users into roles based on where they are in that space (c.f. Matei, in press).

There is some evidence that people do genealogy in different ways, with different goals, and we hope that clustering-based role assignment will bring out some of these differences. For example, most genealogists research their own direct ancestors, searching for proof of dates, places, and relationships. Others concentrate their research on a single place (e.g., everyone who lived in a given town) or on those with a certain surname. Another group is derisively called "name collectors" or "name gatherers". These genealogists are seen as being willing to establish links between people without sufficient evidence, and genealogy forums are replete with complaints about unsourced or poorly sourced information that is shared online (Willever-Farr & Forte, 2014). Clustering may allow us to identify users who fill these sorts of roles.

We begin by removing all of the user-months in which the total number of edits is less than 5. These edits make up our first behavioral role – "Low Activity". We then use machine learning techniques similar to those used by Chan et al. (2010) to automatically identify user roles based on behavior. Specifically, we first rescale all of the features to have similar magnitude, using the "scale" function in R. When

using many clustering algorithms, you must specify the number of clusters (k) that you want the algorithm to produce. We first computed k-medoid clusters for two to ten clusters, using the "pam" function in the R package "cluster", and then chose the number of clusters with the highest average silhouette width (Maechler, Rousseeuw, Struyf, Hubert, & Hornik, 2014). A k of three provided the highest average silhouette width, although the value (0.364) was still very low, and suggests that these clusters may not be a good way of partitioning the behavioral roles.

To test for robustness, we also used the k-means machine learning algorithm. For k-means, there is not a simple measure like silhouette width to compare. In cases like this, when the number of clusters is unknown, one can graph the percent of variance explained for each k, and then choose the "elbow point" – the point at which the amount of variance explained by adding a new k begins to plateau. In this case, our knowledge of the community allows us to look at both the elbow point and the attributes of the clusters that are returned, and to decide which k is most reasonable based on whether additional clusters provide intuitively different groups.

Again, the data do not provide a clear elbow point, so we keep the number of clusters (3) provided by the "pam" function. Once the behavioral roles are identified, we use the mean and median behavioral measures to describe each of the roles. These measures allow us to answer RQ1, whether some of the clustered roles will be identifiable as full membership in a community of practice.

It is important to note that although the actual clustering algorithm we used is completely unsupervised and undirected, this process is semi-inductive. The features that we chose to include, and especially the way that we combined page types into larger categories (such as "local talk" and "community talk") were chosen based on an *a priori* assumption that there would be a core role, and that members of that role would differ along those dimensions. In addition, our interpretation of why roles differ along the various categories is also subjective, and influenced by our research, as well as experience in the community.

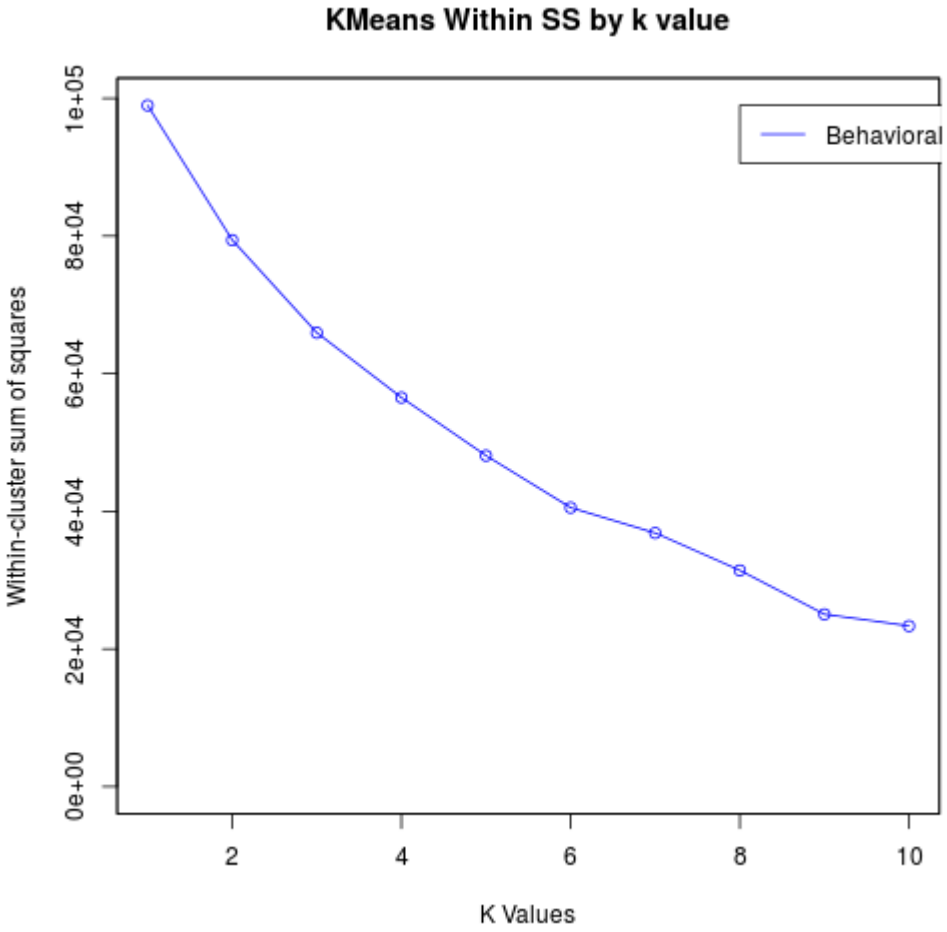

Fig. 3.4: K-means results. Amount of variance in behavior explained by each number of clusters chosen.

3.3.1 Network Positions of Roles

For RQ2, and H1 through H3a, we examine the network measures, (which we refer to as the "network signature") for each of the 161 users, in each of the interaction networks (Stark & Vedres, 2006, cf.). For each 30-day period, we dichotomize each interaction network at two interactions. That is, a tie exists between users only if the weight of that tie is at least two. Because we want to get an overall picture of each node's place in these networks, we then combine all eight snapshots of each network, and dichotomize again. In the end, for each interaction network, we have a network

where a tie represents i and j interacting at least two times in one or more periods. For each user, we record the behavioral role that they were in most often during their entire time on the site, and consider each user as belonging to that role.

We examine some of the measures which paint a picture of the nature of a user's relationship with the network – we measure the degree, the local clustering coefficient, and the eigenvector centrality for each node. Degree is simply a measure of the number of other nodes one is connected to. We treat this as an undirected measure, even if the graph is directed. The local clustering coefficient measures how connected one's neighbors are to each other. For example, if I communicate with A, B, and C, and A communicates with B and C, but B and C don't communicate with each other, then my local clustering coefficient is 2/3, since 2/3 of the possible connections between my communication partners exist (Watts & Strogatz, 1998). Eigenvector centrality is a measure of a node's "influence" in a network, and takes into account both the number of nodes one is connected to, and the centrality of the nodes one is connected to (Hanneman & Riddle, 2005). We compare the mean and median network statistics for each role, for each of our interaction networks.

For RQ2, we look at the general distribution of network statistics. For H1 and H2, we look at the eigenvector centrality scores, predicting that the core membership role will have the highest scores, particularly in the community talk network. For H3a, we look at the differences between clustering coefficient scores for the different roles.

For H3b, we compare the way that users are clustered into behavioral roles with the way they are clustered into network communities. We use the Walktrap random walk community-detection algorithm, as implemented in the R igraph package, to detect communities based on ties, for each of the combined networks (Pons & Latapy, 2005). These communities are visualized via images of each network, with nodes colored by modal role and background colors based on the Walktrap community they are a part of. We also look at the proportion of role members that are in the same community. We identify the Walktrap community that has the highest number of members of each role, for each network. We then look at the proportions of members of that role

which are in that top community. As a hypothetical example, if the core role had ten members, and the local communication network had three Walktrap communities, we would find the community that had the most core role members, and divide that number by the total number of core role members. If there were seven members in the top community, for example, then the proportion would be 0.7.

3.3.2 Role Transitions

To examine the pathways of role evolution (H4a, H4b, and H5), we provide summary statistics, graphical representations, and statistical analyses. These analyses are something of a simplified version of the methods used by Matei et al. (2014) to study the dynamic behaviors of active Wikipedians. First, we simply look at the roles that precede each other. We include "Quitting" as the "role" for the period following the final period in which a user was active on the site. We look at the overall and conditional ratios for the roles that users move into from other roles.

For the graphical representations, we compare the pathways that individual users take, starting with the period in which they made their first edit (for a similar analysis, see Panzarasa, Opsahl, & Carley, 2009). We show a summary of which role each user participated in, for each month of their tenure on the site. For example, the first measurement includes the role that each user had in the first month that they were active (and not the first month that the site existed).

We create two types of graphs. For the first graph, we show the number of users in each role, starting in their first month of membership at the left. For the second graph, we show the ratio of users in each role, again with the first month they made an edit on the left. As users stop editing, they are removed from the graph, so the fifth month, for example, includes the fifth month for only those users who were still active on the site five months after they joined. Therefore, the ratios will always add up to 1, even though the number of users on the right side of the graph will be very small.

Finally, we run two regression analyses. To examine whether early behavioral roles predict future behavioral roles, we look at the first three months of activity for all users. We create a binary dependent variable based on whether or not a user spends at least two months (after the first three) as a member of the core role. We then use the roles a user is in for their first three months as predictors. If different roles are good predictors in early months than late months, then we could take this as evidence that users follow a pathway toward full membership in the community of practice.

To examine quitting the community, we chose a random month from the last ten months of the community to use as our focus month. This is within a time period where the site had the highest activity, and is recent enough to allow us to contact users involved, if needed. The period chosen was the 30 days ending 3 July 2012. We took all of the users who made at least one edit during the previous month, and determined whether or not they quit during the focus month. We defined quitting as making less than two edits during the focus month (this number was chosen to ensure that all crosstabs included at least 5 observations). We then looked at how the roles that a user was in for the three months previous to the focus month affected whether or not they quit during the focus month.

3.4 Stochastic Actor-Oriented Models

For the remaining hypotheses, we use the RSiena stochastic actor-oriented model (SAOM) simulation software to model the behavior of users over time. Because networks are inherently interdependent, traditional statistical methods are not applicable. It is particularly difficult to determine whether behavior is a result of social influence or homophily. For example, trying to determine whether someone drinks alcohol because their friends drink, or if they choose to be friends with people who also drink. Indeed, there is an argument that these effects are impossible to completely disentangle (Shalizi & Thomas, 2011). Stochastic actor-oriented modeling is

specifically designed to try to get around this problem. RSiena simulates behavioral and tie changes in a longitudinal network based on parameters of interest (such as social influence), while controlling for other parameters, such as in-degree, out-degree, homophily, and cyclicality. It compares these stochastically generated networks to the observed network, and iteratively updates the parameters to improve the fit. When the process is over, we are left with a set of parameters and standard errors, allowing for significance testing of the parameters (Steglich, Snijders, & Pearson, 2010).

Because this process is stochastic, and relies on computationally intensive modeling, it is advisable to start with a simple model, and iteratively add and remove effects to find a model that converges, and that addresses the questions of interest (Ripley, Snijders, Boda, Vörös, & Preciado, n.d.). Instead of trying to fit multiple questions into one model, we ran simpler RSiena models for each of our hypotheses. This has the downside of not controlling for potentially important effects, but allows for models to converge, and to run in a much shorter time. For models that we could not get to converge, we try to control for time heterogeneity as outlined by Ripley et al. (n.d.). For some models, we also changed the level of dichotomization, so that a tie exists only when there were three interactions in a month, rather than two. When even that is not successful, we apply the modeling to a subset of the networks. The changes made are noted in the Results section.

3.4.1 Behavior Contagion

For H6a-H6c, we ran a simple RSiena model for each network, which includes rate constants for each period (not shown in the tables), some controls for homophily, transitivity, graph density, and controlling for how long a user has been in the community (based on quintile).

3.4.2 Centrality and Full Membership

For each of the remaining hypotheses, we model a network that includes all interactions. We create this network similar to the way that we combined networks across time for earlier analyses. We dichotomize each of the different interaction networks for a given month, such that a tie exists if there are at least two interactions of the same type between two users. We then sum all of the resultant matrices, and dichotomize again, such that a tie exists in the combined network if it exists in any of the interaction networks.

For H7, we examine whether high eigenvector centrality predicts joining the core role. To do this, we use RSiena's "creation function". This function looks only at those who do not currently have a behavior (like being in the Core Member role), and uses given effects to predict whether a given user will add that behavior in the next period (Ripley et al., n.d.). As before, we control for density, transitivity, and balance, as well as the general tendency to be in the core role.

3.4.3 Tie Decay

For H8a and H8b, we look at why ties between users decay. While the creation function focuses only on behaviors (or network ties) that don't currently exist, RSiena has an "endowment function", which focuses only on behaviors or ties that do exist, and modeled effects predict whether a tie which currently exists is likely to remain (Ripley et al., n.d.). We begin with a model which controls for network density, reciprocity, and balance, and then add measures for similarity between users for each of the roles. For H8b, we include selected interactions between roles, to see if there are any complementary roles. We began with a model that included a large number of interactions, and removed those which were least theoretically significant in order to find a better fit.

3.4.4 Quitting the Community

For H9a and H9b, we create a single RSiena model to test the effects of eigenvector centrality and embeddedness on quitting the community. We use eigenvector centrality as defined previously. For embeddedness, we are trying to get at the notion that those who are part of a large, tight-knit group are less likely to quit. We therefore define embeddedness as the number of alters multiplied by the local clustering coefficient (i.e., the ratio of your alters who are connected to each other).

Because actually quitting the community was not modeled in the data used to create the networks, we use moving into the Low Activity role as a proxy for quitting. This role includes all users who made less than five edits on the site in a month, and will include those actually quit the community. We control for density, reciprocity, transitivity, balance, general tendency to quit, and whether a user is in the core role.

4. RESULTS

4.1 Role Descriptions

To answer our first research question, we identify whether any clusters/roles bear the hallmarks of communities of practice. Namely, they should be active in the community, should communicate with other users, and should participate in activities that require training and acculturation, such as the "complex" pages.

Table 4.1: Mean and median values of clustered behavioral roles

Measure	Core Members	Peripheral Experts	Newbies	Low Activity
User-months	2612	1930	6455	33955
Unique users	507	1110	4014	6249
Mean (Median)				
Active days	17.613	4.207	2.057	0.237
	(17.000)	(2.000)	(1.000)	(0.000)
Days since first edit	592.034	377.556	271.041	505.858
	(422.000)	(60.000)	(29.000)	(375.00)
Simple total	0.807	0.181	0.826	0.087
	(0.914)	(0.044)	(0.990)	(0.000)
Community talk total	0.008	0.022	0.004	0.003
	(0.000)	(0.000)	(0.000)	(0.000)
Community total	0.006	0.016	0.002	0.001
	(0.000)	(0.000)	(0.000)	(0.000)
Complex total	0.110	0.584	0.042	0.068
	(0.050)	(0.571)	(0.000)	(0.000)
Other total	0.021	0.028	0.006	0.005
	(0.000)	(0.000)	(0.000)	(0.000)
Local talk total	0.047	0.103	0.020	0.051
	(0.006)	(0.000)	(0.000)	(0.000)
Logged edits	6.926	4.549	3.695	0.223
	(6.707)	(4.174)	(3.497)	(0.000)
Logged manual edits	6.156	2.691	2.516	0.200
	(6.010)	(2.485)	(2.565)	(0.000)

We show the mean and median values for each of the behavioral measures in Table 4.1. The "user-months" refers to the number of observations that were put into each role (where an observation is one month of activity for one user). Unique users is the number of users who had at least one month in a given role.

The first role, which we call "Core Members", has the characteristics of core members of a community of practice, providing an affirmative answer to RQ1. They are by far the most active role, both in the number of days per month that they make edits, and in the number of edits that they make. They have also been members of the site much longer than the other roles and participate more in talk (although even for this group, talk edits make up less than 1 percent of edits).

We call the second role "Peripheral Experts". While these members are active for only a few days per month, they make a large number of edits, and a surprisingly large proportion of edits in the "Complex" category. We looked at the original data, and found that many of these edits occur in the "Place" and "Image" categories. Also interesting is the fact that they make a very large number of automated edits (total edits - manual edits). This group may represent users who do most of their family history work on a local program on their computer, and then periodically upload GEDCOM files to update WeRelate. Because we do not include GEDCOM uploads in our analysis of edit ratios, this explains why they would appear to have fewer edits in simple namespaces.

The third role we call "Newbies", which contains the majority of the clustered user-months. These users are the least active (with a median of only 1 day of editing per month). They are also the role that has most recently joined the community. Nearly all of their edits occur in the simple namespaces. Overall, this role appears to be composed almost completely of the initial activity as users join the site.

Finally, for reference we include the "Low Activity" role. This role was created manually, and is composed of all of the user-months in which the user made less than five edits. The "Days since first edit" is larger than we might expect. This may be an artifact of how we created our data. While we remove all user-months following a

user's last edit, there are a number of users who were active for a while, then inactive for a long period of time, and then came back to the site to make another edit. All of the months of inactivity are recorded as Low Activity months, and skew the "Days since first edit" value.

4.2 Role-based Network Measures

We next examine summary network measures for each of the identified roles (Table 4.2). For these measures, we only examine the 161 active users identified for RSiena modeling, and we use their modal role, over their entire tenure on the site.

We see that, at least for this group of users, there appear to be two distinct groups. The Core Members and Peripheral Experts have comparatively high degree, high centrality, and high clustering coefficients for all four networks. The Low Activity and Newbie members have similarly low degree, centrality, and clustering coefficients for each network.

These results provide partial support for H1, H2, and H3a. H1 and H2 predicted that the core members would have the highest eigenvector centrality scores. While their scores are much higher than those of typical members, the Peripheral Experts have similarly high scores. In fact, in the Community Talk network, where we predicted core members would have particularly high centrality, they actually have lower centrality than Peripheral Experts. H3a predicted that the clustering coefficient would also be highest for core members. Again, we see that clustering coefficients are higher for core members than for typical members, for all networks. However, the Peripheral Experts actually have higher clustering in every network. We decided not to treat the Peripheral Expert role as a core membership role, but these results provide some evidence that these users are very important to the interaction networks on the site.

H3b suggested that not only would interaction networks show core members to be in tight-knit groups, but that many of the other members of those groups would also be core members. In Figures 4.1 through 4.4, we show visualizations of each

Table 4.2: Mean and median network statistics for each role

Measure	Core Members	Low Activity	Newbies	Peripheral Experts
Number of Users	53	76	26	6
Mean (Median)				
Observation Degree	22.849	3.276	2.462	20.333
	(13.000)	(1.000)	(1.500)	(20.000)
Local Talk Degree	9.585	0.895	0.769	7.333
	(4.000)	(0.000)	(0.000)	(7.000)
Community Talk Degree	9.132	1.684	2.538	14.333
	(0.000)	(0.000)	(0.000)	(7.000)
Collaboration Degree	6.453	0.842	0.769	7.000
	(2.000)	(0.000)	(0.000)	(5.000)
Observation EigCent	0.289	0.057	0.044	0.292
	(0.196)	(0.020)	(0.021)	(0.304)
Local Talk EigCent	0.206	0.026	0.016	0.190
	(0.087)	(0.000)	(0.000)	(0.200)
Community Talk EigCent	0.216	0.043	0.058	0.340
	(0.000)	(0.000)	(0.000)	(0.153)
Collaboration EigCent	0.159	0.026	0.020	0.235
	(0.079)	(0.000)	(0.000)	(0.226)
Observation LocalClustCoef	0.320	0.182	0.250	0.430
	(0.252)	(0.000)	(0.000)	(0.298)
Local Talk LocalCC	0.048	0.015	0.008	0.089
	(0.037)	(0.000)	(0.000)	(0.116)
Community Talk LocalCC	0.058	0.015	0.011	0.121
	(0.000)	(0.000)	(0.000)	(0.141)
Collaboration LocalCC	0.032	0.006	0.009	0.082
	(0.000)	(0.000)	(0.000)	(0.095)

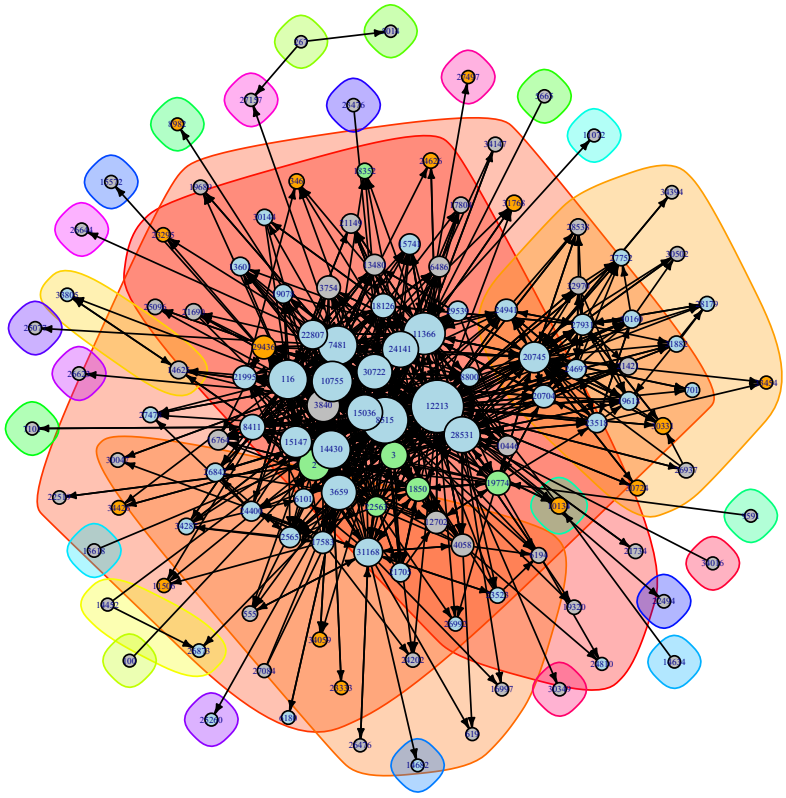

Fig. 4.1: Graph of Walktrap-based communities in observation network, with nodes colored by behavioral role and sized by eigenvector centrality. Newbies are orange, Low Activity members are gray, Core Members are blue, and Peripheral Experts are green. Background colors are arbitrary, and distinguish Walktrap-based communities.

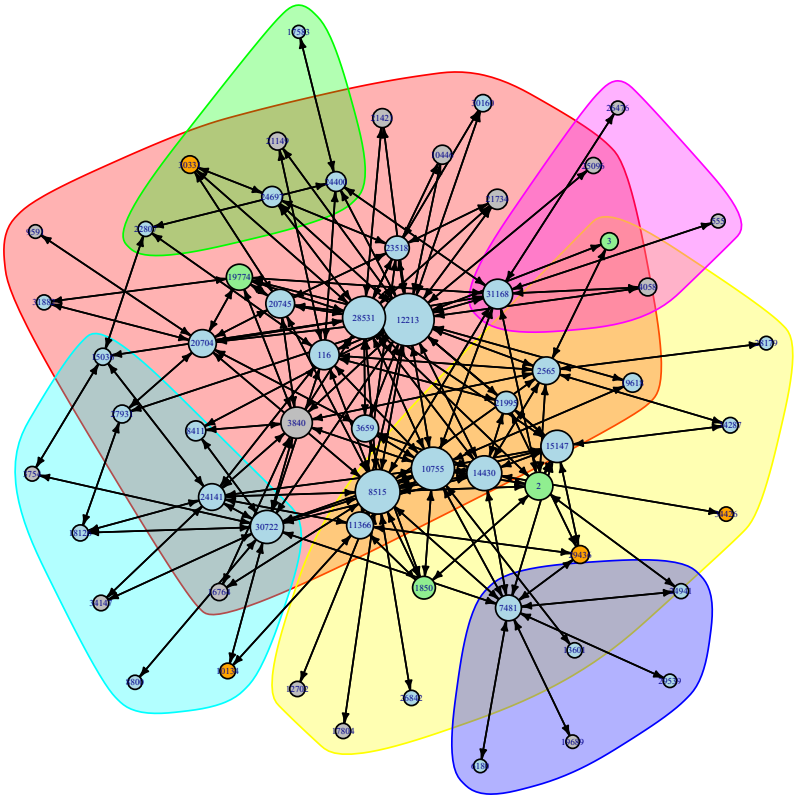

Fig. 4.2: Graph of Walktrap-based communities in local communication network, with nodes colored by behavioral role, and sized by eigenvector centrality.

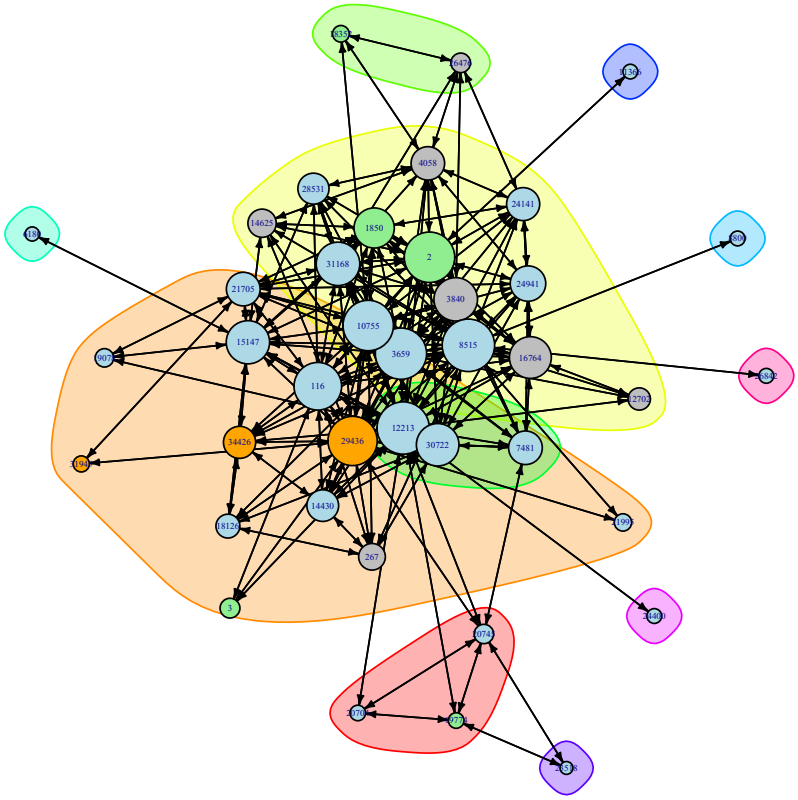

Fig. 4.3: Graph of Walktrap-based communities in community-wide communication network, with nodes colored by behavioral role, and sized by eigenvector centrality.

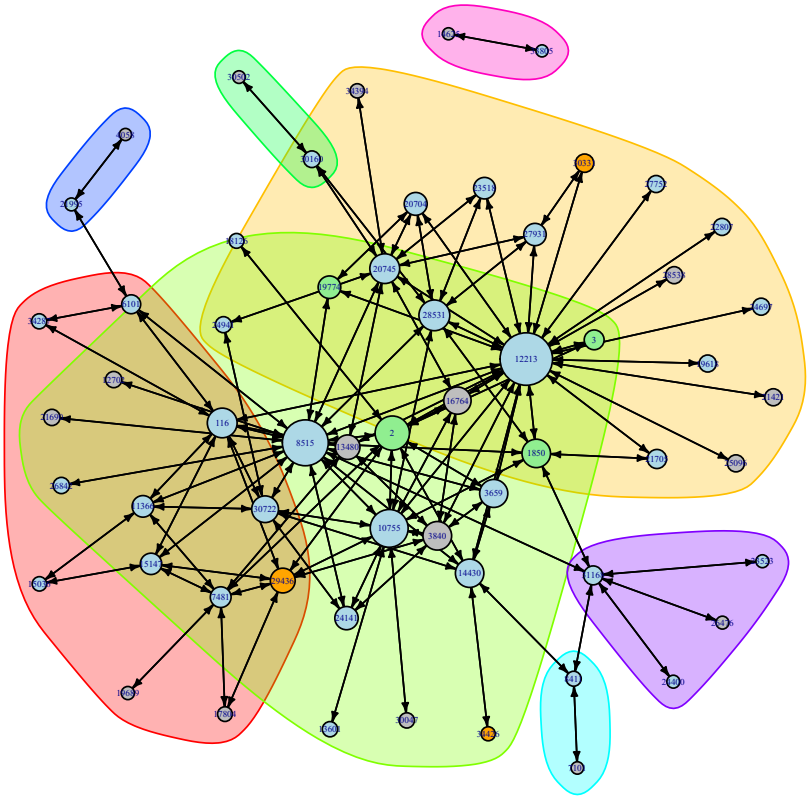

Fig. 4.4: Graph of Walktrap-based communities in collaboration network, with nodes colored by behavioral role, and sized by eigenvector centrality.

of the eight-month combined networks. For each network, the nodes are sized by eigenvector centrality. Newbies are orange, Low Activity members are gray, Core Members are blue, and Peripheral Experts are green. Polygons are drawn around each of the network-based communities created by the Walktrap algorithm, with a different color for each network-based community. Isolates have been removed.

When we look visually at these networks, we see that Core Members really do make up the backbone of each of these networks, occupying the center of the graph in every case. In addition, they are generally positioned near each other.

The observation network is by far the most dense. This is not surprising, since nearly all edits contribute to this network, while other networks have more stringent requirements. The few Peripheral Experts (green nodes) appear to be less important in the observation and local communication networks, but they have quite high centrality in the community-wide communication network, perhaps indicating that this may be where they share their expertise. In addition, a few Newbies (orange nodes) are quite central in the community-wide communication network. These may be new users who are using the community-wide help pages to seek advice or help in using the site.

H3b suggested that these network-based communities would be homogeneous. In Table 4.3, we look at the Walktrap-based community that has the most members of each role, and determine what proportion of that role's members is in that community. We find that the Core Members are very clustered together, with a much higher ratio of role members all in the same Walktrap-based community than for Low Activity or Newbies, in every network. However, as before, these ratios are comparable to (and are in fact lower than) the ratios for the Peripheral Expert role. Again, this provides partial support for H3b, since the Core Members are much more likely to be in the same Walktrap community than typical members, but are not more likely to be in the same community than the Peripheral Experts.

Table 4.3: Proportion of role members in one network-based community

	Observation	Local Talk	Community Talk	Collaboration
Core Members	0.566	0.283	0.170	0.208
Low Activity	0.236	0.083	0.056	0.083
Newbies	0.345	0.103	0.034	0.034
Peripheral Experts	0.714	0.286	0.286	0.286

4.3 Role Pathways

We next look at how members move from one role to another, over time. We begin with Table 4.4, which shows each of the possible role transitions. There is a count of how many times a role transition occurred, the percent of total role transitions that this represents, and finally, a percentage of role transitions, conditioned on the earlier role. For example, 0.47% of all transitions are from Core Member to Low Activity, but this represents 8.24% of all transitions from Core Member to anything.

Table 4.4: Role transition statistics

Previous Role		Current Role	Count	Percent of Total	Percent of Previous Role
Low Activity	→	Low Activity	26336	58.98%	77.81%
Low Activity	→	Core Member	174	0.39%	0.51%
Low Activity	→	Peripheral Expert	418	0.94%	1.24%
Low Activity	→	Newbie	1635	3.66%	4.83%
Low Activity	→	Quitting	5282	11.83%	15.61%
Core Member	→	Low Activity	210	0.47%	8.24%
Core Member	→	Core Member	1710	3.83%	67.11%
Core Member	→	Peripheral Expert	136	0.30%	5.34%
Core Member	→	Newbie	415	0.93%	16.29%
Core Member	→	Quitting	77	0.17%	3.02%
Peripheral Expert	→	Low Activity	645	1.44%	33.79%
Peripheral Expert	→	Core Member	130	0.29%	6.81%
Peripheral Expert	→	Peripheral Expert	388	0.87%	20.32%
Peripheral Expert	→	Newbie	144	0.32%	7.54%
Peripheral Expert	→	Quitting	602	1.35%	31.53%
Newbie	→	Low Activity	1755	3.93%	27.62%
Newbie	→	Core Member	298	0.67%	4.69%
Newbie	→	Peripheral Expert	161	0.36%	2.53%
Newbie	→	Newbie	961	2.15%	15.12%
Newbie	→	Quitting	3179	7.12%	50.03%

The majority of people are in the Low Activity role, making fewer than 5 edits per month, and they stay in that role. Fully 59% of all transitions are from one low-activity month to another, meaning that a user hasn't yet quit the community, but they aren't very involved. As before, some of this behavior may be explained by our data collection methods. Also apparent is that Core Members stay in that role, more than others. Periods in which someone is a Core Member are followed by another

period as a Core Member 67% of the time. Core Members are also very unlikely to quit the community or to move into Low Activity (11% combined).

Both Peripheral Experts and Newbies are fairly likely to leave the community (32% and 50%, respectively) or to move into Low Activity (34% and 28%, respectively). They also often stay in their same role (20% and 15%, respectively), but they are not very likely to move into any of the other roles.

4.3.1 Role Transition Visualizations

To test H4a, we look at whether the early behavior of a user predicts their future behavior. We start by examining role transitions graphically. Each set of graphs includes the number of users in each role on the left, and the ratio by role on the right.

We first plot the full frequency of roles over time in Figure 4.5. The most obvious result is what we already know – most users quit very quickly. In this graph, we include the month following a user's final edit in the Low Activity role. Because most users quit in their first month, there is a huge jump in Low Activity in the second period, indicating that many people edit the first period, then never again.

In Figure 4.6, we show a stacked graph of the ratios of members in each role at each time period. We still see that most people (from about 55% to 85% in any given month) are in the Low Activity role. We also see that the distribution of roles is quite stable for a long period of time. However, after being in the community for about three years, there does appear to be a general upward trend in the other roles, suggesting that people who remain in the community for that long become more active (or, alternatively, that people who are more active remain in the community).

To try to tease out the difference between these two, we examine a number of different graphs based on the role that each person is in most often. Starting with the users whose modal role is Low Activity, we see that most of them quit the community very early (Figure 4.7). By the third period, most users have left the community.

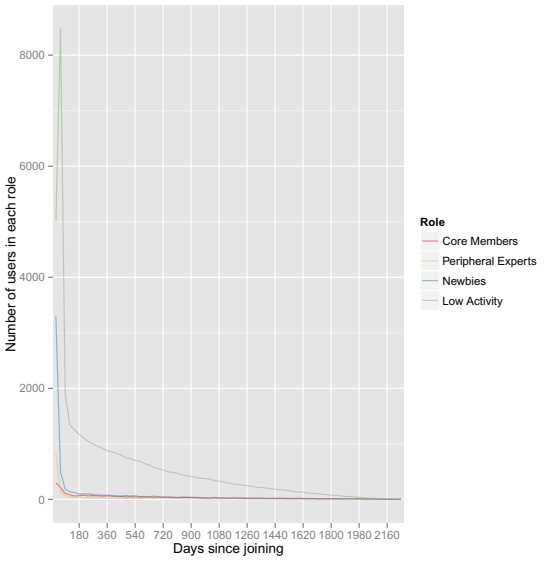

Fig. 4.5: Number of users in each role, by days since their first edit.

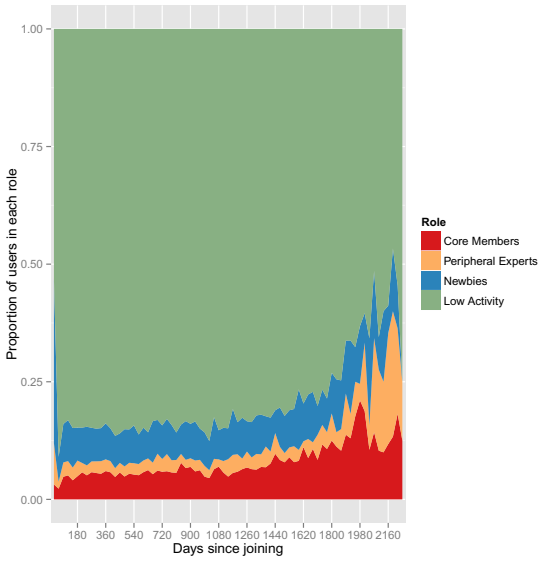

Fig. 4.6: Ratio of users in each role, by days since their first edit.

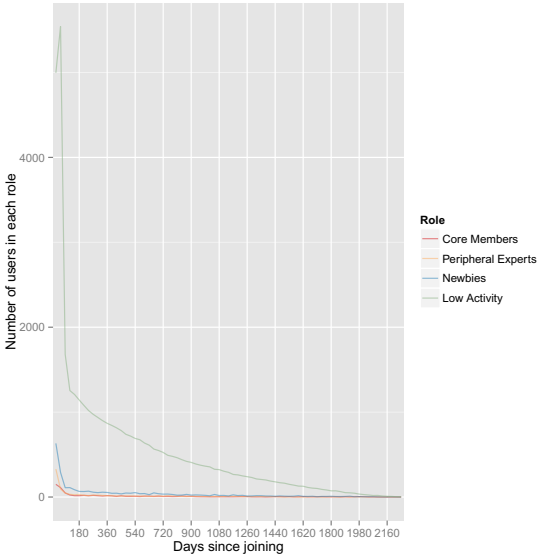

Fig. 4.7: Number of modal Low Activity users in each role, by days since first edit.

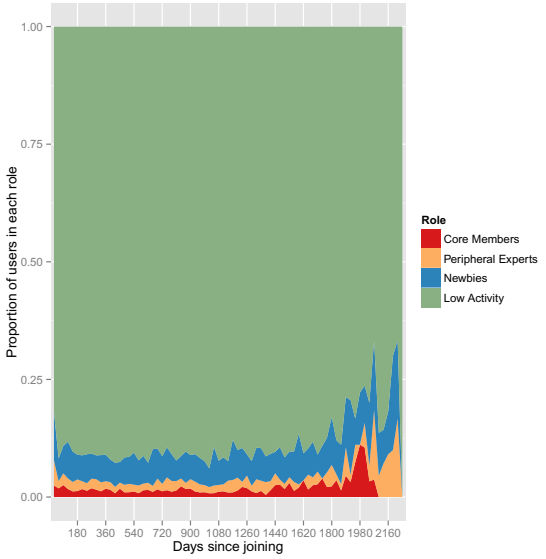

Fig. 4.8: Ratio of modal Low Activity users in each role, by days since first edit.

Indeed, even when we look at the ratio of users by role, there isn't much to suggest that users are likely to join the more active roles as they stay in the community longer (Figure 4.8). Interestingly, there is a fairly large group of people who remain connected to the community for a long time, despite generally contributing very, very little. As mentioned earlier, this may simply represent users who are active very sporadically – making edits, then leaving the community for a long time before contributing again.

Those whose modal role was a Newbie had an even steeper drop off (Figure 4.9). Hardly anyone is left within a few months. This suggests that this role – almost exclusively editing simple pages – is something that only new users do, and they then leave the community. Because nearly everyone is gone, the ratio graph (Figure 4.10) is not meaningful after the first few months.

The Peripheral Expert graph suggests that this role is also composed almost entirely of new users, as the drop off within the first few periods is extreme (Figure 4.11). When we look at the ratio of users in each role, there is some evidence that there is a group of users who are persistently Peripheral Experts, but that group is obviously quite small (Figure 4.12).

As suggested by the transition table, the only group that seems to have real staying power is the Core Members. A huge proportion of them start out in the Core Member role in their very first month, and they stay there (Figures 4.13 and 4.14). However, there is no clear evidence of a progression of roles from peripheral to core membership.

4.3.2 Regression Analysis

We also ran a binomial logistic regression model (Table 4.5), which gives a hint that there may be something going on which is not visible in the graphical representations, i.e., that there may be a pattern followed by those who eventually become Core Members. Even from their first month, those who are in the Core Member and

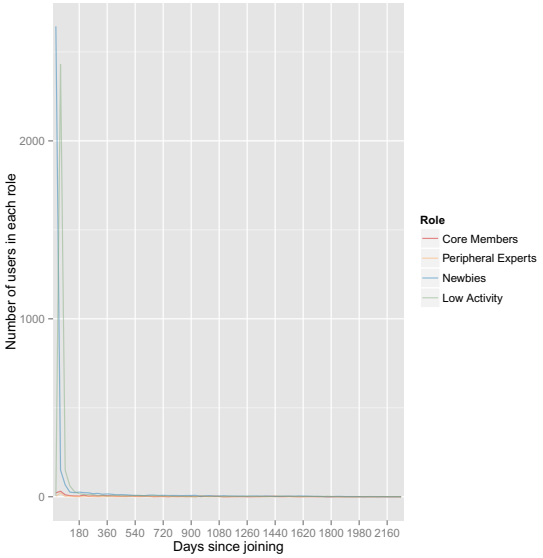

Fig. 4.9: Number of modal Newbie users in each role, by days since first edit.

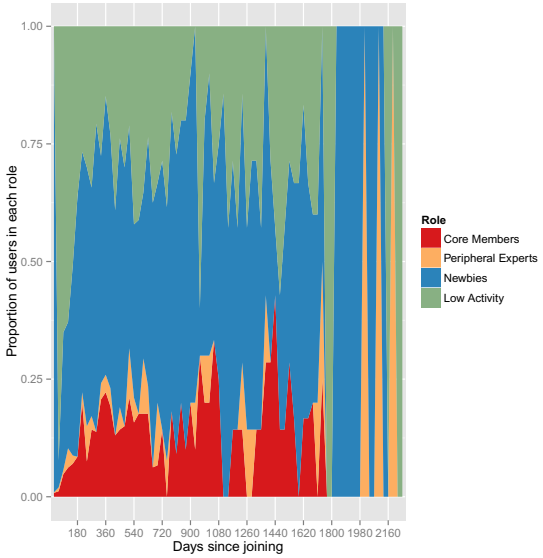

Fig. 4.10: Ratio of modal Newbie users in each role, by days since first edit.

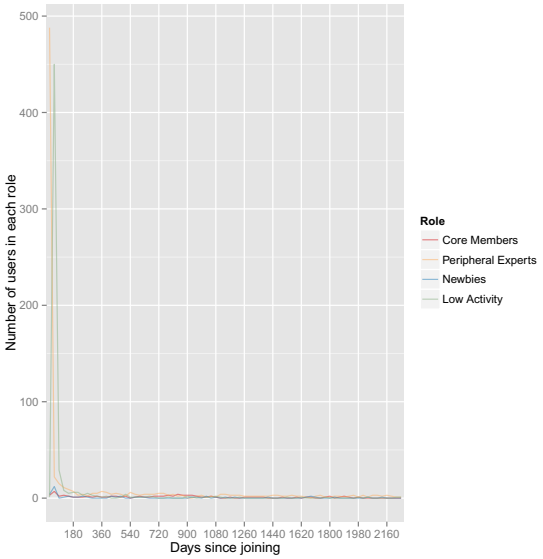

Fig. 4.11: Number of modal Peripheral Expert users in each role, by days since first edit.

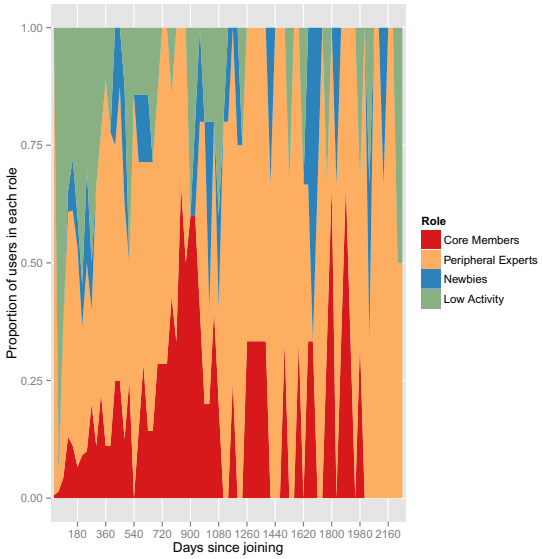

Fig. 4.12: Ratio of modal Peripheral Expert users in each role, by days since first edit.

Fig. 4.13: Number of modal Core Member users in each role, by days since first edit.

Fig. 4.14: Ratio of modal Core Member users in each role, by days since first edit

Table 4.5: Logistic regression of early roles as predictors of future core roles

	Dependent variable:
	Became Core Member
T1 CoreMember	0.923^{***}
	(0.328)
T1 Newbie	0.376
	(0.303)
T1 PeripheralExpert	0.651^{*}
	(0.339)
T2 CoreMember	2.002^{***}
	(0.305)
T2 Newbie	1.137^{***}
	(0.283)
T2 PeripheralExpert	0.381
	(0.498)
T3 CoreMember	1.631^{***}
	(0.318)
T3 Newbie	0.152
	(0.303)
T3 PeripheralExpert	1.056^{***}
	(0.381)
Constant	-3.820^{***}
	(0.277)
Observations	1,613
Log Likelihood	−374.763
Akaike Inf. Crit.	769.525

Note: T1 represents a user's role during their first month in the community, T2 the role in the second month, and T3 their role in their third month.

*p<0.1; **p<0.05; ***p<0.01

Peripheral Expert roles are more likely to eventually be in the Core Member role, controlling for the roles they are in for their second and third months. However, this relationship increases in strength for both the second and third months. In the first month, those who are in the Core Member role are 2.5 times more likely than those in the Low Activity role to be in the Core Member role in the future ($p = .005$) In the second and third months, this jumps up to 7.4 and 5.1 times more likely, respectively ($p < .001$). That is, acting like a Core Member in your second or third month is a better predictor of your future behavior than the role you are in for your first month. This result is tempered by noting that the way we gathered data means that users may have joined the site partway through their first month, and this may provide a partial explanation for why roles in the first month are less useful predictors.

Taken together, these results provide strong support for H4a. Those who become Core Members begin their time on the site acting differently than typical members. They are much more likely to be in some of the more active roles.

On the other hand, these results provide mixed and modest support for H4b. There is no clear path from initiate to Core Member, but there is some evidence that later roles are a better predictor of core membership than earlier roles, which implies some sort of initiation or learning.

4.3.3 Analysis of Quitting Behavior

In order to examine how people leave the community (H5), we flip things around. We identify those who have quit as users who had no activity in the final month of measurements, and then find the last month that they had any activity. We then plot all of the roles that users were in, with their last month of activity at the left end of the graph. That is, instead of synchronizing users based on when they made their first edit, we synchronize those who quit, based on when they made their last edit. Moving left to right, therefore, goes backward in time. We follow a similar

methodology as before, showing all users, then showing graphs for the users whose modal role was each of the behavioral roles.

When we look at all users (Figure 4.15), we don't see anything too surprising. Most people quit very quickly, and most of them are in the Low Activity and Newbie roles. As we move backward in time, the ratio of Core Members and Peripheral Experts increases, while the absolute count does not, indicating that those who take core roles simply stay in the community for longer (Figure 4.16).

Looking at the Low Activity users (Figures 4.17 and 4.18), we find a very similar pattern, reflecting the fact that most members of the community are Low Activity users, and then quit. This also indicates that there are a number of users who start off and remain Low Activity users for a significant period of time before quitting.

Looking at the Newbie leavers (Figures 4.19 and 4.20), we see that the vast, vast majority quit after their first 30 day period. There are so few long-term users who are modal Newbies, that the ratio of users in each role is fairly meaningless.

The Peripheral Expert plot is a little more interesting. Again, most of these users leave immediately after their first edits (Figure 4.21). However, for those who stay, there is a fairly clear pattern. Looking at Figure 4.22, as we approach the time of departure from the right, we see an increasing proportion of Low Activity roles, indicating that these users begin to disengage from the community during the last months before they leave.

The plot of Core Members tells a similar story (Figures 4.23 and 4.24). As we near the time of departure (left side of the graph), the ratio of Newbie and Low Activity behavior increases quite dramatically, while Core Member activity decreases. These plots appear to indicate that while there is not a clear path into full activity in this community, departing full activity is more predictable and visible.

A logistic regression predicting extremely little to no activity provides further evidence (Table 4.6). The results show that users who were Core Members in the previous month have a 95% decrease in their likelihood of quitting, compared to those in the Low Activity role ($p < .001$). Those in the Peripheral Expert and Newbie roles

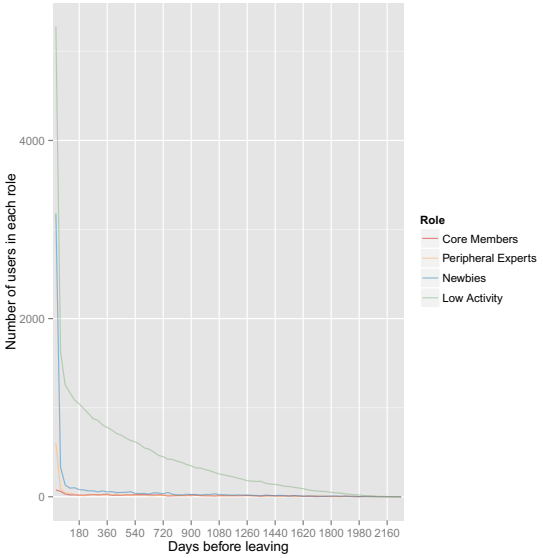

Fig. 4.15: All leavers: Number of users in each role, by days before final edit.

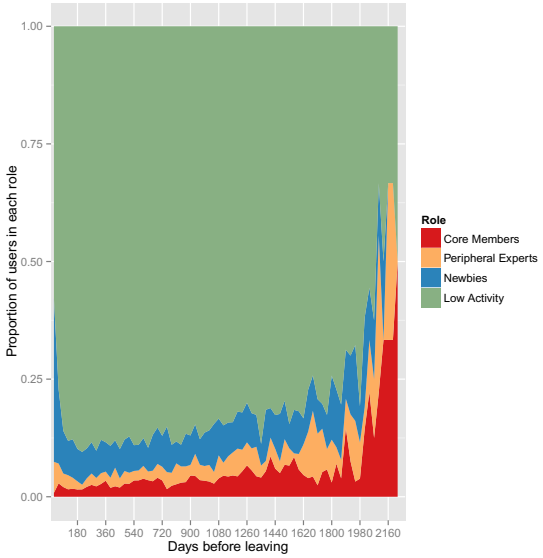

Fig. 4.16: All leavers: Ratio of users in each role, by days before final edit.

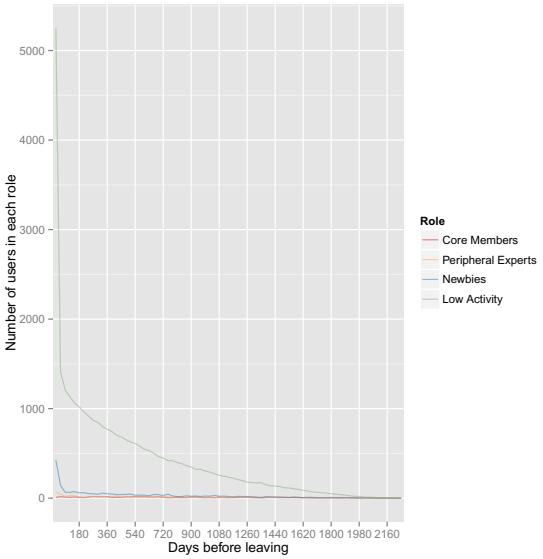

Fig. 4.17: Modal Low Activity leavers: Number of users in each role, by days before final edit.

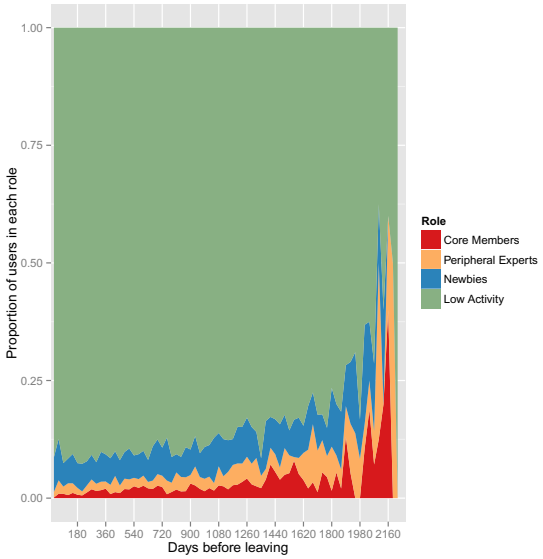

Fig. 4.18: Modal Low Activity leavers: Ratio of users in each role, by days before final edit.

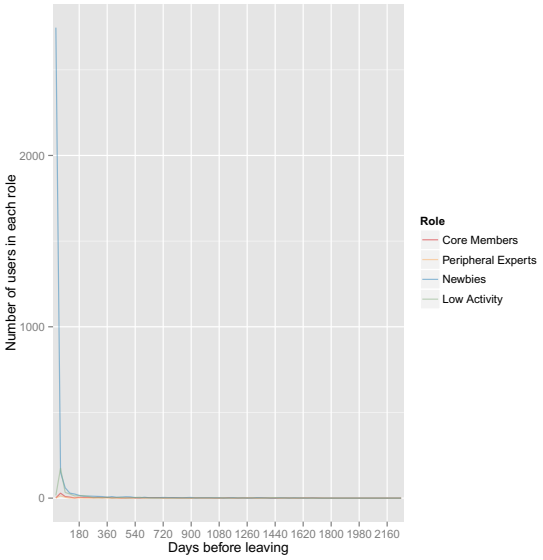

Fig. 4.19: Modal Newbie leavers: Number of users in each role, by days before final edit.

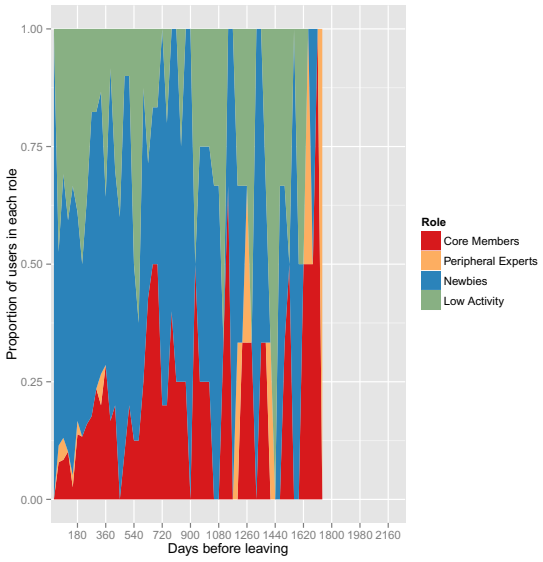

Fig. 4.20: Modal Newbie leavers: Ratio of users in each role, by days before final edit.

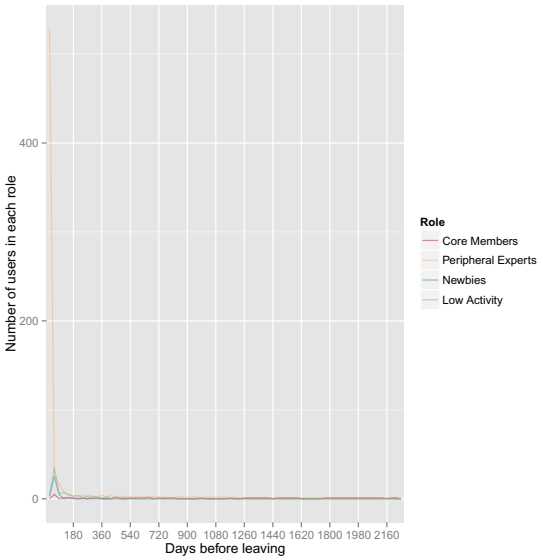

Fig. 4.21: Modal Peripheral Expert leavers: Number of users in each role, by days before final edit.

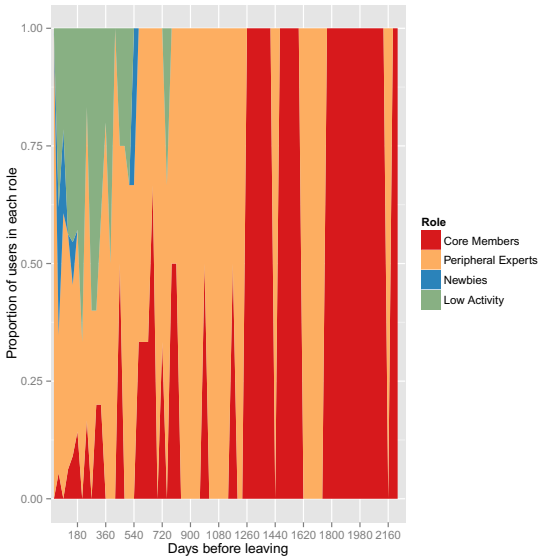

Fig. 4.22: Modal Peripheral Expert leavers: Ratio of users in each role, by days before final edit.

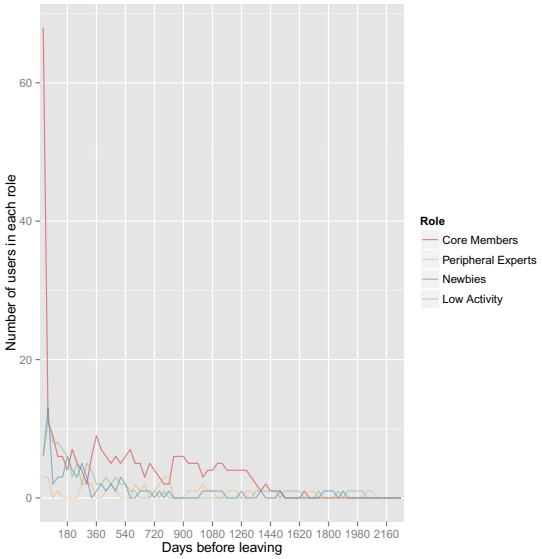

Fig. 4.23: Modal Core Member leavers: Number of users in each role, by days before final edit.

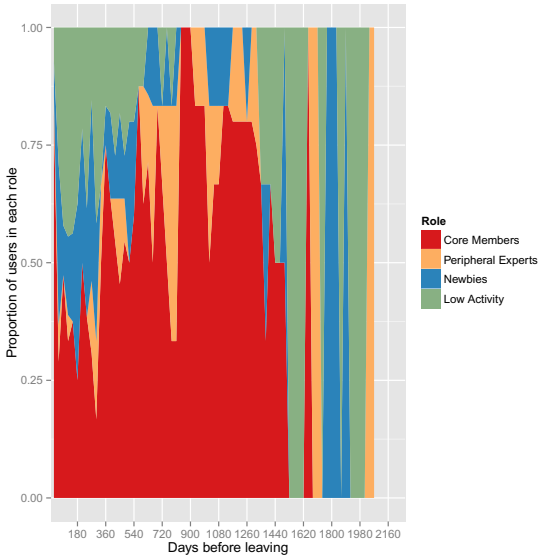

Fig. 4.24: Modal Core Member leavers: Ratio of users in each role, by days before final edit.

Table 4.6: Logistic regression predicting whether a user will quit contributing

	Dependent variable:
	Stopped Contributing
OneBeforeCoreMember	-2.921^{***}
	(0.574)
OneBeforeNewbie	-1.707^{***}
	(0.311)
OneBeforePeripheralExpert	-2.827^{***}
	(0.655)
TwoBeforeCoreMember	-1.514^{**}
	(0.664)
TwoBeforeNewbie	-0.857^{**}
	(0.412)
TwoBeforePeripheralExpert	-2.299^{***}
	(0.642)
TwoBeforeNotYetJoined	-1.315
	(1.149)
ThreeBeforeCoreMember	-0.998^{*}
	(0.585)
ThreeBeforeNewbie	-0.716^{*}
	(0.420)
ThreeBeforePeripheralExpert	-0.671
	(0.675)
ThreeBeforeNotYetJoined	2.486^{**}
	(1.110)
daysSinceJoining	-0.0002
	(0.0002)
Constant	2.653^{***}
	(0.304)
Observations	628
Log Likelihood	-201.921
Akaike Inf. Crit.	429.842

Note: OneBefore represents the user's role in the month before their final edit, TwoBefore their role two months before, etc.

*p<0.1; **p<0.05; ***p<0.01

were 94% and 82% less likely to quit, respectively ($p < .001$). As we move further back in time, the effect and significance of these roles becomes less pronounced. At two months back, those in the Core Member role, Peripheral Expert role, and Newbie role are 78%, 90%, and 58% less likely to quit, respectively ($p < .05$, $p < .001$, $p < .05$). None of the behavior roles three months before are statistically different from Low Activity, providing evidence that Core Members transition to less active roles previous to quitting.

4.4 Network and Behavior Interactions

For our later hypotheses, we use RSiena's stochastic actor-oriented modeling to study the interaction of behavior and networks. We begin by examining the way that each interaction network affects whether a user joins/stays in the core role. We only find a model that fits moderately well for the observation network (convergence t-ratios all $< .15$; Table 4.7). We find evidence that when one observes those who are currently in the core role, they are more likely to stay in or move into the core role themselves, controlling for network tendencies toward reciprocity, transitivity, and homophily, as well as general likelihood for moving into the core role ($p < .05$). We could not get a model that converged well for the observation network until we changed the dichotomization to three interactions, and this table reports results from that level.

For each of the other networks, convergence was poor (at least 1 convergence t-ratio $> .15$; Tables 4.8 to 4.11). In these other networks, our predictor of interest (behavior CoreMember average alter) is significant in only the combined network (Table 4.11), although it is in the right direction in all networks. In the case of community communication, one of the observation periods made convergence impossible, so we tested only a subset of the observations.

Because the observation network provided a successful convergence with a significant effect for alters influencing behavior, we ran one additional model to dig a little

Table 4.7: RSiena results for observation network

Effect	par.	(s.e.)
Network Dynamics		
outdegree (density)	–4.118***	(0.105)
reciprocity	2.874***	(0.136)
transitive triplets	0.628***	(0.033)
CoreMember similarity	–0.549***	(0.115)
Behaviour Dynamics		
behavior CoreMember linear shape	–2.234**	(0.684)
behavior CoreMember average alter	7.833*	(3.172)
behavior CoreMember: effect from daysSinceJoining	0.128	(0.282)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.145,
overall maximum convergence ratio 0.286.

Table 4.8: RSiena results for local communication network

Effect	par.	(s.e.)
Network Dynamics		
degree (density)	–5.826***	(0.176)
transitive triads	2.249***	(0.278)
balance	–0.199**	(0.064)
CoreMember similarity	0.510	(0.357)
daysSinceJoining similarity	0.243	(0.314)
Dummy7:localCom ego	0.660	(0.720)
Behaviour Dynamics		
behavior CoreMember linear shape	–1.262	(0.879)
behavior CoreMember average alter	51.912	(6219.803)
behavior CoreMember: effect from daysSinceJoining	0.080	(0.229)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.654,
overall maximum convergence ratio 1.097.

Table 4.9: RSiena results for community communication network

Effect	par.	(s.e.)
Network Dynamics		
degree (density)	–5.556***	(0.261)
transitive triads	1.010***	(0.057)
balance	–0.131***	(0.036)
CoreMember similarity	–0.280	(0.562)
daysSinceJoining similarity	0.512	(0.506)
Behaviour Dynamics		
behavior CoreMember linear shape	–0.984*	(0.407)
behavior CoreMember average alter	4.391	(6.900)
behavior CoreMember: effect from daysSinceJoining	0.101	(0.233)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.214,
overall maximum convergence ratio 0.373.

Table 4.10: RSiena results for collaboration network

Effect	par.	(s.e.)
Network Dynamics		
degree (density)	–6.863***	(0.408)
transitive triads	4.229***	(0.727)
balance	–0.438**	(0.146)
CoreMember similarity	1.305	(1.004)
daysSinceJoining similarity	–0.284	(0.489)
Behaviour Dynamics		
behavior CoreMember linear shape	–1.058***	(0.156)
behavior CoreMember average alter	7.690	(11.466)
behavior CoreMember: effect from daysSinceJoining	0.097	(0.140)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.301,
overall maximum convergence ratio 0.686.

Table 4.11: RSiena results for combined network

Effect	par.	(s.e.)
Network Dynamics		
outdegree (density)	–4.231***	(0.087)
reciprocity	3.827***	(0.108)
balance	–0.066***	(0.004)
CoreMember similarity	0.236*	(0.104)
daysSinceJoining similarity	–0.368*	(0.151)
daysSinceJoining ego x daysSinceJoining alter	0.098***	(0.021)
Behaviour Dynamics		
behavior CoreMember linear shape	–2.374**	(0.734)
behavior CoreMember average alter	7.655*	(3.071)
behavior CoreMember: effect from daysSinceJoining	–0.018	(0.263)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.194,
overall maximum convergence ratio 0.601.

Table 4.12: RSiena results for observation network with creation function

Effect	par.	(s.e.)
Network Dynamics		
outdegree (density)	–4.093***	(0.096)
reciprocity	2.811***	(0.136)
transitive triplets	0.622***	(0.031)
coreMember similarity	–0.611***	(0.146)
Behaviour Dynamics		
behavior coreMember linear shape	–1.722*	(0.698)
behavior coreMember average alter	7.385**	(2.743)
inc. beh. coreMember average alter	–3.865	(5.115)
behavior coreMember: effect from daysSinceJoining	0.426	(0.534)
inc. beh. coreMember: effect from daysSinceJoining	–0.464	(0.635)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.265,
overall maximum convergence ratio 0.718.

deeper (Table 4.12). The default behavior in RSiena is to try to predict whether a behavior (or tie) will exist in the following period, independent of whether it exists now. This is called an "evaluation function". As explained in the Methods chapter, RSiena does allow for both "creation" and "endowment" functions, which test for whether a behavior is likely to be adopted, or to be maintained, respectively. We added a "creation" term to the analysis, which was negative (although n.s.). This actually indicates that the effect of observing the work of someone in the core role is more important for users that are already in the core role than it is for users who are in other roles. Thus, we find some very limited support for H6a, but do not find support for H6b or H6c.

We next look at the effect of centrality on becoming a Core Member (Table 4.13). We were unable to find a model that converged well for the full set of networks, but a model that includes only the first four networks did converge successfully (convergence t-ratios $< .1$) We tested whether users who were not in the Core Member role were more likely to become part of the core role in the next period. This predictor (inc. beh. coreRole: effect from eigenvectorCentrality) was positive, indicating that having

Table 4.13: RSiena results for effect of eigenvector centrality on moving into the Core Member role

Effect	par.	(s.e.)
Network Dynamics		
outdegree (density)	–3.909***	(0.103)
reciprocity	2.838***	(0.153)
transitive triplets	0.259***	(0.015)
balance	–0.044***	(0.003)
Behaviour Dynamics		
behavior coreMember linear shape	–0.135	(0.467)
inc. beh. coreMember: effect from eigenvectorCentrality	55.396*	(26.512)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.078,
overall maximum convergence ratio 0.271.

higher eigenvector centrality does increase one's likelihood to become a Core Member ($p < .05$). This result provides support for H7.

We next look at the results of an RSiena model predicting which interaction ties between members are likely to decay. We are able to successfully fit a model (Table 4.14) that looks at how homogeneity between roles affects whether a tie is likely to decay (convergence t-ratios $< .15$). However, homogeneity is not significant for any of the roles, and we do not have enough evidence to support H8a.

It was difficult to fit a model that included interactions between roles (Table 4.15). By simplifying the model, and only using the first four periods, we were able to find a model that converged very well (convergence t-ratios $< .1$). As with the model for role homogeneity, none of the effects are significant, and this model does not provide support for H8b.

We finally examine the effect of embeddedness and eigenvector centrality on quitting the community (Table 4.16). As with some other analyses, we were unable to fit a good model using the full set of networks. A model that includes only the first four networks did converge successfully (convergence t-ratios $< .15$) This model provides evidence that those with high eigenvector centrality are less likely to move into the Low Activity role ($p < .05$), controlling for whether a user is in the core role, as well as their embeddedness. Embeddedness in the community does not provide a significant effect on moving into the Low Activity role ($\alpha = .05$). This model provides some support for H9b, but does not support H9a.

Table 4.14: RSiena results for the effect of role homogeneity on tie maintenance

Effect	par.	(s.e.)
outdegree (density)	–3.723***	(0.052)
reciprocity	3.470***	(0.077)
balance	–0.049***	(0.003)
lowActivityRole similarity	–0.201	(0.311)
coreMember similarity	–0.085	(0.217)
periphExpertRole similarity	0.224	(0.227)
newbieRole similarity	0.398†	(0.240)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.135,
overall maximum convergence ratio 0.255.

Table 4.15: RSiena results for the effect of role complementarity on tie maintenance

Effect	par.	(s.e.)
outdegree (density)	–3.910***	(0.095)
reciprocity	3.407***	(0.131)
balance	–0.053***	(0.003)
coreMember similarity	–0.230	(0.225)
periphExpertRole similarity	0.031	(0.543)
periphExpertRole ego x coreMember alter	–0.628	(1.007)
coreMember ego x periphExpertRole alter	0.655	(0.941)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.064,
overall maximum convergence ratio 0.126.

Table 4.16: RSiena results for the effect of embeddedness and centrality on quitting

Effect	par.	(s.e.)
Network Dynamics		
outdegree (density)	–3.888***	(0.101)
reciprocity	2.771***	(0.141)
transitive triplets	0.260***	(0.015)
balance	–0.044***	(0.003)
Behaviour Dynamics		
behavior LowActivity linear shape	–1.106***	(0.227)
behavior LowActivity: effect from CoreMember	–2.197***	(0.513)
behavior LowActivity: effect from eigenvectorCentrality	–19.024*	(7.606)
behavior LowActivity: effect from embeddedness	0.222	(0.259)

† $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$;
convergence t ratios all < 0.114,
overall maximum convergence ratio 0.466.

Table 4.17: Summary of Results

Research Question / Hypothesis	**Results**
ROLES	
RQ1. Will some behavioral roles fit the profile of core membership?	Affirmative
RQ2. Will behavioral roles have unique network signatures?	Mostly affirmative
ROLES AND NETWORKS	
H1. Members of roles representing core members in the community of practice will have higher eigenvector centrality scores than those in peripheral roles, for all interaction networks.	Partially supported
H2. The difference in eigenvector centrality between core members and others will be greatest in the community talk network.	Partially supported
H3a. Core members will be in more tightly clustered groups than others, for all interaction networks.	Partially supported
H3b. Core members will be more likely than other roles to be in the same networkÂŋbased clusters as each other.	Partially supported
LONGITUDINAL ROLE EVOLUTION	
H4a. Users who become core participants will participate in different behavioral roles than others from the very beginning of their time on the site.	Supported
H4b. Users who eventually move into core participant roles will begin their tenure on the site in peripheral roles.	Partially supported
H5. The recent behavioral roles that users have been in will predict whether they will stop participating in the community.	Supported
SOCIAL INFLUENCE	
H6a. Users who observe the work of core participants will be more likely to move into core participant roles.	Partially supported
H6b. Users who collaborate with core participants will be more likely to move into core participant roles.	Not supported
H6c. Users who communicate with core participants will be more likely to move into core participant roles.	Not supported
H7. Users who have high eigenvector centrality in interaction networks while in peripheral roles will be more likely to advance to core participation roles.	Supported
TIE DECAY	
H8a. Ties between members with the same role will be less likely to decay.	Not supported
H8b. Ties between members with different roles will be less likely to decay.	Not supported
H9a. Users who are embedded in interaction networks will be less likely to quit the community.	Not supported
H9b. Users who are central in interaction networks will be less likely to quit the community.	Partially supported

5. DISCUSSION

In this chapter, we put some of the findings of this paper in context, discuss some of the limitations and caveats, and describe some future directions for research that could be interesting.

5.1 Role Clustering

Clustering user behavior into roles was the most exploratory part of our study. We found that WeRelate is similar to other peer production projects in many ways. Like other projects, participation follows a very skewed distribution, with a few people doing a large portion of the work. Also like other projects, most participants come and go very quickly, without contributing much. Our clustering was effective in finding evidence of peripheral and core roles, and clustering provided good evidence that these roles existed. In addition, we were able to confirm, as predicted by the Situated Learning literature, that the core users were both more active and more central in interaction networks.

However, we had also hoped that clustering would illuminate different modes of working on the site that would not be visible to a casual user. While the clustering did appear to effectively identify the core members in the community of practice, other roles were more ambiguous and less helpful.

In addition, measures of fit indicated that this data set may not have been a good candidate for this type of clustering. This problem is compounded because the roles that we created provided an important input to much of the rest of our analysis. When dealing with network analysis on a group of only 161 users, having users misclassified or ambiguously classified makes it difficult to produce good models, and may have contributed to our difficulties with RSiena. However, our main goal in these later

analyses was to identify and compare core and peripheral members, and these did appear to be classified quite well, so the effects of this ambiguity should not have been too severe.

There were a few factors that may have contributed to our difficulty in finding accurate and interesting roles. The sources of data that we have from the website are limited - we know how many edits a user made within each period, and what categories those edits appear in. Some of the more nuanced differences in the way that people use the site may be difficult to distinguish from this sort of data. Also, in this paper we used a very simple clustering algorithm, as a means of showing that this could be an effective way of both identifying key users and of modeling how people move through a community. Allow us to suggest a few ways that our clustering could have been improved:

- Do some clustering on only long-term users, to tease out new modes of working without being overwhelmed by the noise of the majority who quit quickly.
- Do some clustering on only the first few months of activity. This may allow for more fine-grained ways of identifying differences in how users begin their tenure.
- Create labeled data. In some communities, users sign up as members of certain roles. These could be used to train a classifier.
- Include additional features. We focused on how active users were, and which types of pages they edited. Other measures such as the proportion of their edits which create new pages, the number of edit made per editing session, or the length of text per edit could provide useful insights and make clustering more effective.

Finally, it is possible that even with these improvements, clustering would still not be reliable enough to provide robust, machine-only categorization of users. Even in that case, clustering can prove useful to get a "lay of the land" of a community. Clustering can help to identify otherwise invisible modes of behavior for further analysis.

For example, our "Peripheral Expert" role showed us that there are people who are active on the site much less often than "Core Members", and who make fewer edits, but who are still very important in interaction networks. Looking more in-depth at who these users are, and what sorts of edits they make, could prove very useful in understanding the community. This sort of model flips the strategy used by Gleave et al. (2009) and Welser et al. (2011), by beginning with automated classification of roles, which can then be used to inform and guide a more qualitative examination of the community.

We believe that overall, our results show that clustering behavioral roles is a promising approach, with potential for simplifying and illuminating aspects of a community that might otherwise be hidden.

5.2 Role Transitions

As in Wikipedia (Panciera et al., 2009), we found that the early behaviors that a user engages in are very predictive of whether they will eventually become a core member of the community. That is, core members are different from the beginning, and remain different. Our results tell the story of one role (Core Members) whose membership is stable, and is much more likely to remain involved in the community than any of the other roles. The regression analysis that we ran provided some evidence that for at least some members, their first and second months include some learning and changing of behavior, but overall, our results did not provide clear evidence of a path from initiate to peripheral member to core participant. This suggests that, at least for this community, core members are recruited rather than trained and initiated, a concept discussed in further detail in our discussion of the network models.

In perhaps our most interesting finding, we discovered that users have distinctive patterns of behavior before disengaging with the site. This provides evidence that people don't simply quit, but go through a period of deepening withdrawal. This

suggests that there may be opportunities to identify users who are on the verge of quitting, and to intervene before they leave completely. Research into discovering exactly what sorts of changes to behavior are involved with quitting would be both interesting and useful in creating tools to identify these users.

5.3 Networks and Stochastic Modeling

When we added a network dimension to our analysis of role transitions, we were able to shed some additional light on what is happening. We were unable to find strong evidence that interacting with core members has an effect on peripheral members. Overall, there was only slight evidence that users undergo a period of acculturation, learning, or peripheral participation, and only the observation network provided any significant results.

One reason for these results might be that WeRelate is a fairly technical website in a niche where there are a lot of other options which are simpler to use. There may be a selection bias, where those who choose to remain on the site may be those who already have technical and genealogical skills. Therefore, interactions would be less important for learning and support.

One other possible reason is that users of WeRelate work much more independently than many other peer production projects. On Wikipedia, for example, popular articles have thousands of contributions from hundreds of members. Even niche subjects often attract interest from a few dozen people. Genealogists generally work primarily on their own ancestors, and so for many pages, a user may be the only person who ever edits the page. Unlike Wikipedia, there are not a lot of general-interest pages where new users can interact with others, so there may not be many opportunities for norm creation and knowledge transfer.

While new members may not be affected by their relationship with core members, we found that those who are already core members are less likely to leave the core role if they interact with other core members.

These findings suggest a few conclusions about the way this community works. First, core members are fundamentally different from peripheral members. This makes some sense. The dedication required to be a core member requires the sort of time and disposition that many people simply don't have, and we shouldn't expect most people to change easily. Those who come into the project with these attributes will almost certainly display different behavior from the start.

Second, those who are invested in the project have a hard time leaving. They decrease their activity over months before finally giving up completely. This may represent changes in situation (such as getting a different job, having a child, etc.), or changes in one's feelings toward the project. Either way, it seems to show evidence of an attachment to the site that is lasting.

If these conclusions are correct, then it appears that if WeRelate members want to increase the number of core members who participate, they could take the following actions. First, they can work to recruit people who are like them – those who are already active in genealogy and have time on their hands. Second, they can identify the new members who are on the trajectory toward core membership, and make sure they interact with other core members. This can be done by identifying those who are particularly active, or who are in the active roles within their first few months of membership. Finally, they can identify core members who experience a marked reduction in activity, since this is a marker of those who are on the verge of leaving. These users can then be drawn back in through increased interaction with other members.

These findings have theoretical implications, as well as practical. Together with the results from research on Wikipedia, by Panciera et al. (2009), our results suggest that large scale online peer production projects differ from traditional communities of practice in a few ways. Most importantly, for the majority of contributors, these projects are not communities of practice. While offline participation in a group requires a certain level of commitment (finding out when they meet, making time in your schedule, attending the meeting, etc.), online participation requires much less ef-

fort. Users can create an account and make a few edits to the site on a whim, and then never come back, which is exactly what most do. In addition, computer-mediated communication research suggests that the text-based nature of many of these communities makes transmission of norms and knowledge more difficult and time-consuming (Walther, 2011). This may explain why we did not find much evidence of situated learning via communication networks.

On the other hand, our results also show the presence of a definite community of core contributors, who are not only more active, but who interact with each other in the creation of this joint artifact. The creation and sustaining of that sort of community certainly requires some sort of acculturation and coordination, and it seems like some sort of situated learning must be going on, either in ways that our analysis missed, or externally to the site. At any rate, we believe that Situated Learning is a useful way to think about and examine online communities, and using a theory of group formation as a framework helped us to frame and answer questions about how these communities work.

5.3.1 Issues with Stochastic Modeling

Many of the questions we answered using RSiena did not have significant results. Our RSiena models did not converge easily, and required a lot of coaxing. In addition, while convergence is a measure of goodness of fit for the selected effects, RSiena also makes available the option to check a few non-included measures (such as degree distribution). We checked some of these goodness of fit measures for a few of the networks, and while some produced good results, others did not, indicating that the models may not truly be a good representation of how these networks and behaviors evolve.

There are a few reasons that we may have had these difficulties. The most important is that stochastic actor-oriented models are designed to analyze the effect of deeper relationships. While we believe that the different types of networks that we

outlined provide useful distinctions in the types of interactions that people have on WeRelate, they may be too granular for stochastic actor-oriented modeling. Indeed, SAOM may be a poor fit in general for examining the effects of fleeting, short-lived interactions, such as those that occur on WeRelate and many other online collaborative platforms. In their article on RSiena, Snijders, Van de Bunt, and Steglich (2010) say

> A dynamic network consists of ties between actors that change over time. A foundational assumption of the models discussed in this paper is that the network ties are not brief events, but can be regarded as states with a tendency to endure over time. Many relations commonly studied in network analysis naturally satisfy this requirement of gradual change, such as friendship, trust, and cooperation. Other networks more strongly resemble 'event data', e.g., the set of all telephone calls among a group of actors at any given time point, or the set of all e-mails being sent at any given time point. While it is meaningful to interpret these networks as indicators of communication, it is not plausible to treat their ties as enduring states, although it often is possible to aggregate event intensity over a certain period and then view these aggregates as indicators of states.

Our data certainly fall under the description of "event data". We have attempted to aggregate intensity, by only including ties where there were at least two (or sometimes three) events between people in a month. However, at such a low cutoff it is difficult to argue that the ties between users really represent states/relationships. But, moving the cutoff any higher would make what are already very sparse networks even more sparse, and finding any significant results even more difficult. Indeed, many of our questions are about whether new users who have incidental contact with established users are more likely to stick around. We are specifically wondering whether "events" can lead to relationships, and if we chose to focus only on relationships, we would be unable to address these sorts of questions.

While SAOM provide a good statistical foundation for their results, as well as some hints at causality, other sorts of analysis, such as ERGM, do not have the same limitations for tie creation, and might be more appropriate for our data. Another option might be the dynamic matched sample estimation used by Aral et al. (2009), although our dataset might not be large enough.

5.4 Data Collection and Interpretation

In addition to analysis concerns, there are also some general concerns and caveats that apply to many studies like ours. The data available from WeRelate and other digital projects is incredibly granular. We know exactly which page was edited, when (to the minute), and by whom. As discussed in the introduction, this provides an amazing wealth of information, and has ignited entire fields of research. However, there are a number of potential issues for research using digital trace data, particularly in conjunction with networks. Howison et al. (2011) identify ten issues, a number of which may apply to our study.

First, assumptions must be made about the meaning of interactions. This data does not represent survey-based responses to data; rather, researchers must make decisions about what sorts of interactions represent which constructs. If we had surveyed WeRelate users about who they collaborated with over the last month, the network created from their responses might or might not look the same as our network. In our case, we have tried to make reasonable assumptions in the way we create the different types of networks in our study, but these assumptions could have been made differently (e.g., we could have only included content pages for our collaboration network), and different sets of assumptions could lead to very different results.

Second, they warn about dichotomizing data. Many tools for network analysis (including RSiena) rely on binary networks, where edges do not contain weight. This results in a loss of data that can be important. For example, there is a great difference

between a new user who has received two messages from a core member and a new user who has exchanged a dozen messages, but they are treated identically in our study. Researchers who use these tools must choose a way to dichotomize the data, and as we saw when running RSiena models, these choices can affect the results.

Third, they warn about inferring non-links. This is definitely a potential problem for our construct of observation. We make a tie from i to j only when i edits a page. However, there are many opportunities to observe the pages others have edited without editing the page oneself, and we certainly miss a large amount of the observation activity that occurs.

Fourth, they warn about the effect that data aggregation can have. This is particularly relevant for our study. In our case, we chose to create networks based on 30-day windows of observation. We seek to find evidence of learning and changing in user behavior - if that change occurs within days or weeks, then our month-long windows will miss it. Howison et al. (2011) suggest running sensitivity analyses to test whether this is the case, but given the complexity of the process of data aggregation and analysis, we did not take this step.

Finally, they warn of directly importing findings from the research on survey-based social networks (which represent more long-term relationships between nodes) into digital trace data networks. While we have used survey-based findings as a basis for some of our research, for the most part, we believe that our study questions whether these findings apply in our context, rather than taking it for granted.

One final caveat that applies specifically to the context of WeRelate is that there are relationships and learning that take place outside of the website. WeRelate is not one of the larger genealogy websites, and does not include resources for searching for genealogical information, like some of the others. Active members of WeRelate almost certainly use some of the other genealogical websites, and interact with other genealogists. These interactions are very likely affect the way that users choose to interact (or not) on WeRelate, and missing these interactions is an important loss.

5.5 Future Work

There are a few directions that we believe this work could be extended, which would be particularly profitable. First, by focusing on the multiplex nature of these communities. That is, by studying the way that the interaction networks that we identified influence and interact with each other. For example, finding out if users who participate in community talk are also likely to collaborate, or if those who observe others also talk with them, etc.

Second, the interpretations could be validated through interviews with community members. In particular, it would be good to find out whether core members experienced a period of acculturation as described by Bryant et al. (2005), or if they really did come into the community and hit the ground running, as the data appears to indicate. In addition, talking with those who have left the community to find out why they left, and why they left gradually, would be useful.

Third, we would love to see this work extended to other contexts. It is possible that some of our findings are very specific to a genealogical context. For example, users may already have been studying their family's genealogy for some time before editing on WeRelate. Their initial activity would then be composed of transferring work they had already done onto WeRelate. Once that activity was complete, they might not have much reason to remain active on the site. This might explain the relatively high number of users who go from the Core Member role in their first few months to quitting. In other communities, we might therefore expect to see fewer Core Members at the beginning of their tenure.

We would also be very interested to know how well clustering behavioral roles would work in a large and broad community like Wikipedia, and whether those roles would match up with the roles identified by others (e.g., Bryant et al., 2005; Welser et al., 2011).

Finally, experimental research could be done to both validate the results that we have found, and to extend them. For example, new users could be randomly

selected to receive a message (or not) from a core member when they join. Our research suggests that this would not make a difference. On the other hand, our results suggest that "at risk" users who received a message from another core user would be less likely to decrease their activity. Experiments could help to establish both the reality and causality of these findings.

6. CONCLUSION

This thesis provides what we believe to be a novel way of understanding an online peer production community, with some interesting and useful results. Our study finds evidence of a community that is dominated by a small group of insiders, who interact often and do much of the work on the site. We find that these insiders come into the community already acting differently from other users, and that while this community may be a community of practice, there is little evidence that legitimate peripheral participation is an important part of it, or at least that this learning does not take place through interactions on the site. However, these users are connected to each other with ties that do not decay easily, and they have a hard time leaving the community.

Understanding the community in this way provides opportunities for targeted recruitment and engagement, as well as an insight into the power structures and opportunities for change. It also furthers the theory of Situated Learning, providing both a new context for studying this theory, as well as empirical results which confirm some of the predictions made by Situated Learning, while pushing back on others. Our study provides an examination of a web-based community from a communicative, user-centric perspective, adding to the growing literature on the structure and evolution of online communities.

We provide a framework for looking at online collaborative communities, and a number of novel measures which we believe to be useful in similar contexts. First, we introduced the concept of machine learning-generated behavioral roles as a simple way of identifying different types of users, and showed that the roles correlate with a number of interaction network measures. We also introduced a new way of creating interaction networks, and showed that insights can be gained by examining the way users interact in various networks, by using both community detection algorithms and

stochastic actor-oriented modeling. Lastly, we introduced a number of visualizations for examining how users move through roles in a community, and showed how these can profitably be combined with tools like regression analysis.

LIST OF REFERENCES

LIST OF REFERENCES

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology*, *29*(2), 97–119.

Aldrich, H., & Ruef, M. (2006). *Organizations evolving* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544–21549.

Baumer, E. P., Adams, P., Khovanskaya, V. D., Liao, T. C., Smith, M. E., Schwanda Sosik, V., & Williams, K. (2013). Limiting, leaving, and (re)lapsing: an exploration of Facebook non-use practices and experiences. In *Proceedings of the 2013 ACM Annual Conference on Human Factors in computing systems* (pp. 3257–3266).

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. Yale University Press.

Bertino, E., & Matei, S. A. (Eds.). (in press). *Roles, trust, and reputation in social media knowledge markets: Theories and methods*. New York: Springer Verlag.

Brandes, U., Kenis, P., Lerner, J., & van Raaij, D. (2009). Network analysis of collaboration structure in wikipedia. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 731–740).

Britt, B. C. (2011). *System-level motivating factors for collaboration on wikipedia: A longitudinal network analysis* (Unpublished master's thesis). Purdue University.

Britt, B. C. (2014). *Evolution and revolution of organizational configurations on wikipedia: A longitudinal network analysis* (Unpublished doctoral dissertation). Purdue Univeristy.

Bryant, S. L., Forte, A., & Bruckman, A. (2005). Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work* (pp. 1–10).

Burt, R. S. (2000). Decay functions. *Social Networks*, *22*(1), 1–28.

Burt, R. S. (2002). Bridge decay. *Social Networks*, *24*(4), 333–363.

Chan, J., Hayes, C., & Daly, E. M. (2010). Decomposing discussion forums and boards using user roles. *ICWSM*, *10*, 215–218.

Choi, B., Alexander, K., Kraut, R. E., & Levine, J. M. (2010). Socialization tactics in wikipedia and their effects. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (pp. 107–116).

Christakis, N. A., & Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, *357*(4), 370–379.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, S95–S120.

Crowston, K., & Howison, J. (2005). The social structure of free and open source software development. *First Monday*, *10*(2).

Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., & Potts, C. (2013). No country for old members: user lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 307–318).

Fisher, D., Smith, M., & Welser, H. T. (n.d.). You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences.*

Fowler, J. H., Christakis, N. A., Steptoe, & Roux, D. (2009). Dynamic spread of happiness in a large social network: longitudinal analysis of the framingham heart study social network. *BMJ: British Medical Journal*, 23–27.

Gleave, E., Welser, H. T., Lento, T. M., & Smith, M. A. (2009). A conceptual and operational definition of 'social role' in online community. In *HICSS'09. 42nd Hawaii International Conference on System Science* (pp. 1–11).

Goffman, E. (1959). The presentation of self in everyday life.

Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods.* University of California,.

Henning, P. H. (1998). Ways of learning an ethnographic study of the work and situated learning of a group of refrigeration service technicians. *Journal of Contemporary Ethnography*, *27*(1), 85–136.

Hertel, G., Niedner, S., & Herrmann, S. (2003). Motivation of software developers in open source projects: an internet-based survey of contributors to the linux kernel. *Research Policy*, *32*(7), 1159–1177.

Hogg, M. A., & Terry, D. I. (2000). Social identity and self-categorization processes in organizational contexts. *Academy of Management Review*, *25*(1), 121–140.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767 - 797.

Iba, T., Nemoto, K., Peters, B., & Gloor, P. A. (2010). Analyzing the creative editing behavior of wikipedia editors: through dynamic social network analysis. *Procedia - Social and Behavioral Sciences*, *2*(4), 6441–6456.

Keegan, B., Gergle, D., & Contractor, N. (2012). Do editors or articles drive collaboration?: multilevel statistical network analysis of wikipedia coauthorship. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 427–436).

Keegan, B., Gergle, D., & Contractor, N. (2013). Hot off the wiki structures and dynamics of wikipedia's coverage of breaking news events. *American Behavioral Scientist*, *57*(5), 595–622.

Kittur, A., Chi, E., Pendleton, B. A., Suh, B., & Mytkowicz, T. (2007). Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web*, *1*(2), 19.

Lakhani, K. R., & Wolf, R. G. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Perspectives on Free and Open Source Software*, *1*, 3–22.

Laniado, D., & Tasso, R. (2011). Co-authorship 2.0: Patterns of collaboration in wikipedia. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia* (pp. 201–210).

Laniado, D., Tasso, R., Volkovich, Y., & Kaltenbrunner, A. (2011). When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *ICWSM*.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge University Press.

Lee, S., & Monge, P. (2011). The coevolution of multiplex communication networks in organizational communities. *Journal of Communication*, *61*(4), 758–779. doi: 10.1111/j.1460-2466.2011.01566.x/full

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web* (pp. 641–650).

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2014). cluster: Cluster analysis basics and extensions [Computer software manual].

Marsden, P. V. (1990). Network data and measurement. *Annual Review of Sociology*, 435–463.

Matei, S. A. (in press). A social network analysis "practice capital" approach to enhance the C-SPAN archive with meta-communication data to support public affairs debates and data journalism. In R. X. Browning (Ed.), *The C-SPAN archives: An interdisciplinary resource for discovery, learning, and engagement.* West Lafayette, IN: Purdue University Press.

Matei, S. A., Bertino, E., Zhu, M., Liu, C., Si, L., & Britt, B. (in press). A research agenda for the study of entropic social structural evolution, functional roles, adhocratic leadership styles, and credibility in online organizations and knowledge markets. In E. Bertino & S. A. Matei (Eds.), *Roles, trust, and reputation in social media knowledge markets: Theories and methods.* Springer Verlag.

Matei, S. A., Wutao, W., Zhu, M., Britt, B., Bertino, E., & Foote, J. (2014). Adhocratic order and the rule of the 1%.

(Manuscript in preparation.)

Mead, G. (1934). *Mind, self, and society from the standpoint of a social behaviorist.* Chicago.

Merton, R. K. (1957). The role-set: Problems in sociological theory. *British Journal of Sociology*, *8*(2), 106–120.

Monge, P. R., & Contractor, N. S. (2003). *Theories of communication networks.* Oxford University Press New York.

Nov, O. (2007). What motivates wikipedians? *Communications of the ACM*, *50*(11), 60–64.

Panciera, K., Halfaker, A., & Terveen, L. (2009). Wikipedians are born, not made: a study of power editors on wikipedia. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work* (pp. 51–60).

Panzarasa, P., Opsahl, T., & Carley, K. M. (2009). Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, *60*(5), 911–932. doi: 10.1002/asi.21015/full

Pons, P., & Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005* (pp. 284–293). Springer.

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction*, *1*(1), 13–32.

Quass, D. (2014). Personal communication.

Reagle, J. (2012, October). 410 gone - infocide in open content communities. *Selected Papers of Internet Research*, *13*.

Ripley, R. M., Snijders, T. A., Boda, Z., Vörös, A., & Preciado, P. (n.d.). Manual for RSIENA [Computer software manual].

Schroer, J., & Hertel, G. (2009). Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, *12*(1), 96–120.

Seabrook, J. (1998). *Deeper: My two year odyssey in cyberspace.* Simon & Schuster Trade.

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239.

Shaw, A., & Hill, B. M. (2014). Laboratories of oligarchy? how the iron law extends to peer production. *Journal of Communication*, *64*(2), 215–238.

Smith, M., Rainie, L., Shneiderman, B., & Himelboim, I. (2014). Mapping twitter topic networks: From polarized crowds to community clusters. *Pew Research Internet Project*.

Snijders, T. A., Van de Bunt, G. G., & Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, *32*(1), 44–60.

Stark, D., & Vedres, B. (2006). Social times of network spaces: Network sequences and foreign investment in hungary1. *American Journal of Sociology*, *111*(5), 1367–1411.

Steglich, C., Snijders, T. A., & Pearson, M. (2010). Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, *40*(1), 329–393. doi: 10.1111/j.1467-9531.2010.01225.x/full

Swartz, A. (2006). Raw thought: Who writes wikipedia. *Blog article available at http://www.aaronsw.com/weblog/whowriteswikipedia Published*, *4*.

Voss, J. (2005). Measuring wikipedia. In *International Conference of the International Society for Scientometrics and Informatics.*

Walther, J. B. (2011). Theories of computer-mediated communication and interpersonal relations. *The handbook of interpersonal communication*, 443–479.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (M. Granovetter, Ed.). Cambridge University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in wikipedia. In *Proceedings of the 2011 iConference* (pp. 122–129).

Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of Social Structure*, *8*(2), 1–32.

Wenger, E. (1998). Communities of practice: Learning as a social system. *Systems Thinker*, *9*(5), 2–3.

Willever-Farr, H. L., & Forte, A. (2014). Family matters: control and conflict in online family history production. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing* (pp. 475–486).

Wood, D. J., & Gray, B. (1991). Toward a comprehensive theory of collaboration. *The Journal of Applied Behavioral Science*, *27*(2), 139–162.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678.

Yakel, E. (2004). Seeking information, seeking connections, seeking meaning: Genealogists and family historians. *Information Research*, *10*(1), 10–1.