Living in the Present: Understanding Long-Term Content Referencing in Enterprise Online Communities

RYAN COMPTON, University of California, Santa Cruz, USA
JEFF WARSHAW, University of California, Santa Cruz, USA
HERNAN BADENES, IBM, Argentina
BARTON SMITH, IBM Research - Almaden, USA
STEVE WHITTAKER, University of California, Santa Cruz, USA

Successful online communities accumulate large amounts of long-term content. However there has been little quantitative, theoretically-motivated exploration of how communities organize such content, nor which community members take responsibility for active organization. We examine one aspect of long-term content organization through link behavior, also exploring role differences between enterprise community leaders and members in the context of life-cycle community models. We first classify how content is linked within posts, identifying usage patterns that organize information within and outside communities. We next present an exploratory quantitative analysis of 2,010 communities including 428,476 posts and 1,246,570 links. We show paradoxically that although mature communities accumulate substantial content, organizing that content using links decreases over time. Further analyses suggest that this arises from a recency bias, with communities being focused on current content. Our results also challenge descriptive lifecycle community models, which propose that regular community members adopt greater responsibility over time. We explore explanations for our findings and implications including new tools that encourage responsibility for active organization, as well as methods for members to revisit critical content.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; Blogs; Wikis;

Additional Key Words and Phrases: Online communities, long-term content management, linking, enterprise, hyperlinks, lifecycle models, roles.

ACM Reference Format:
Ryan Compton, Jeff Warshaw, Hernan Badenes, Barton Smith, and Steve Whittaker. 2018. Living in the Present: Understanding Long-Term Content Referencing in Enterprise Online Communities. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 39 (November 2018), 21 pages. https://doi.org/10.1145/3274308

1 INTRODUCTION

Online communities are now successfully deployed in many contexts including the open internet, social media, and enterprises. They allow participants to distribute knowledge, share expertise, answer questions, or provide social support [39, 41]. In many cases, these online communities have generated extensive shared resources and content. Understanding the long-term practices

Authors' addresses: Ryan Compton, University of California, Santa Cruz, Santa Cruz, CA, 95060, USA, rcompton@ucsc.edu; Jeff Warshaw, University of California, Santa Cruz, Santa Cruz, CA, 95060, USA, jeffreywarshaw@gmail.com; Hernan Badenes, IBM, Buenos Aires, Argentina, herchu@gmail.com; Barton Smith, IBM Research - Almaden, San Jose, California, USA, barton.smith@yahoo.com; Steve Whittaker, University of California, Santa Cruz, Santa Cruz, CA, 95060, USA, swhittak@ucsc.edu.

2573-0142/2018/11-ART39 $15.00
https://doi.org/10.1145/3274308

of these successful communities is critical to inform the design of effective tools, develop success metrics and guidelines for effective online community building. In particular, we need systematic understanding of how long-term communities actively manage ever-growing amounts of content [23, 48]. Effective content management has been argued to be critical for retaining members and ultimate community success, as it provides members with straightforward ways to access important community content [31, 36, 49, 53].

This paper uses quantitative methods to examine two aspects of long-term content management in online enterprise communities by assessing the use of hyperlinks to reference and organize content. First, we explore how communities actively organize their resources by measuring content referencing as content grows. We address the following research questions about whether and how content referencing behaviors change over time as content accumulates: How is long-term content structured through referencing so that newcomers can make sense of accumulated content? How do content creation and referencing coexist and change over time? Second, we examine roles: who takes responsibility for content referencing as communities evolve? It is well known that work is shared unequally in online communities [9, 45, 59], but we know little about how workloads shift among members in the long-term. Do the people who originally created the community remain responsible for content referencing or do newer members adopt responsibility as their level of participation increases [34]?

1.1 Content Management through Referencing

Creation involves generating new community content, and much is already known about creation practices [9, 26, 29, 32, 36]. We define content management as the active organization and annotation of pre-existing content created by the community or by others external to the community, in order to facilitate access to that content. With a few exceptions [50], there has been little empirical research into long-term content management. Understanding management is crucial as successful communities increasingly accumulate content that needs to be referenced for new or returning users. Many complex acts of content management are hard to operationalize across large datasets, because they require manual content analysis that relies on detailed domain knowledge [20, 33]. Our main goal in the current paper is to assess representative content management behaviors across a large dataset, so we chose to explore one specific quantifiable form of content management: hyperlink referencing. Hyperlinks are an efficient and pervasive general method to structure complex information [21], e.g. by referencing underlying content using a structured list or creating readable annotations [42]. Hyperlinks are used in many content management systems, such as wikis, as they promote straightforward access to complex content [14, 42]. Hyperlinks help manage content by creating navigational infrastructure [14], as well as supporting curation [19]. Because hyperlinks do not cover all aspects of content management, we refer to the organizational use of links as referencing and limit the discussion of our results to address these specific referencing styles of content management. Although referencing is pervasive in online collaborative content management tools such as wikis, it is also used in many other tools. We therefore explore referencing via links across a range of social media tools, including wikis, forums and blogs.

1.2 Changes in Roles

Our second research question asks which community members take responsibility for online content referencing. Online community research has shown that a small active subset of users contribute the majority of work in online communities [9, 59], both in creating content and successfully coordinating the work of others [39, 62]. Models that describe long-term community lifecycles have suggested there are shifting responsibilities with leaders and members dominating at different points in the community's evolution [23, 49]. These lifecycle models stress the importance of early

proactive leadership to seed interesting content, set policy, welcome newcomers, etc. [31, 39]. But these models also argue for the importance of apprenticeship so that as the community matures and accumulates content, some regular members gradually assume more responsibility [48, 58]. Such members begin as peripheral participants who simply read and lurk, but over time they take on increasing responsibility with a subset organizing and managing their community. Despite clear theoretical consensus of lifecycle models around apprenticeship and increasing member participation [23, 48] , we are unaware of systematic quantitative analyses of how roles and responsibilities change for long-term content referencing.

This paper therefore examines both how long-term communities actively manage content using referencing links as well as role changes in such referencing over time. To address these questions for enterprise communities, we first explore content management using referencing links at a post level, then determine how accurately lifecycle theories predict role changes compared with actual practice. We use quantitative analyses to address these research questions in the context of mature enterprise communities that have access to a range of social media tools.

1.2.1 Contributions. We quantitatively characterize one important aspect of content management using reference links, where we examine both long-term changes and role differences. We contribute to existing literature by exploring the following questions: First, how does referencing using hyperlinks change over time and how does this relate to content creation? Are links more prevalent over time as content builds up and does accumulation of content lead to increased referencing? Second, as content accumulates, who takes responsibility for referencing: members or leaders? Do members assume more responsibility for referencing over time? Our exploratory findings are counterintuitive. First, active content referencing does not increase as content accumulates and second, contradicting lifecycle models, members never assume full responsibility for referencing. Content analysis suggests that recency bias is a possible reason for the absence of such referencing. We suggest new tools and community building practices that better support content management taking these findings into account.

2 RELATED WORK

2.1 Long-term Community Models

Online communities change over time and different long-term models have been proposed to capture such changes. Wenger et al. [58] describe a sequence of stages: potential, coalescing, maturing, stewardship, and transformation. Communities typically start as loose social networks with the potential for becoming connected. As connections form they coalesce into a community. This matures as members take charge to manage increased content and emerging practices. Finally, a community stagnates as it becomes irrelevant or members go on to other activities. Iriberri and Leroy [23] review multiple prior models of community development proposing a similar 5-stage model to Wenger et al. [58]. They also argue for distinct stages: inception, creation, growth, maturity, and death. Kraut and Resnick [32] also recognize that communities have different needs at different times. They do not explicitly outline a long-term stage model but like Iriberri and Leroy and Wenger et al. [23, 58] they characterize problems confronting communities at different stages. Problems include: startup, attracting and socializing newcomers, encouraging commitment, encouraging contribution, and regulating behavior.

These prior models typically focus on content creation rather than how content is managed over the long-term [38, 49, 54]. Content creation is typically associated with people's internal goals or common interests within the community, for example content is created as people seek information in advice communities, collaborate on a shared software task or solicit emotional support in help communities [37]. Content creation has been analyzed in multiple contexts, including online

advice/support [5, 23, 32], knowledge generation in Wikipedia [45, 56, 62], software development [55], and decomposable tasks [25]. Creation has also been studied across different datatypes including text, software and videos [11, 20, 32, 51]. Most creation models observe a disparity between leader and member roles, with leaders creating substantially more content [9, 44, 45, 49, 62]. To address this inequality, other work characterizes leader strategies to promote greater member involvement in content creation [32, 53], and we discuss these in more detail below.

Community models typically assume that content management activities such as referencing increase as the community matures and content becomes increasingly demanding to organize [23, 48, 49, 53]. Nevertheless, existing long-term models do not specify exactly how community information organization changes as content accumulates, as more participants contribute content, and topics potentially become more disparate.

Despite their plausibility, another issue with long-term models is that they are primarily descriptive; they do not propose measurable criteria for key behaviors defining different lifecycle stages. One important contribution of this paper is therefore to operationalize and test these descriptive models at scale, through a measurable aspect of content management. Specifically, we test whether referencing activities increase as content accumulates, enabling us to characterize how referencing changes over time.

2.2 Community Roles: Differences Between Members and Leaders

We have already mentioned the well-documented observation that not all members contribute equally to their community. Much focus has been on the importance of community leaders for creating content [9, 39, 45], but there has been much less research about how other roles change over a community's lifetime.

2.2.1 Leader Behavior. Prior work discusses how leaders promote successful communities by enacting successful activities to encourage [32, 48], contribute, and read content [9, 22], organize and curate content [48, 58], deal with disruptive behavior [32], create a positive environment [9, 22], foster connections [32, 58], manage new members [32], advertise externally [9], maintain infrastructure [48, 58] and promote longer community life spans [31]. Johnson et al. [24] developed a quantitative model of leader behaviors, in which a leader is defined as having a formal authority role, a highly connected network position, and using leadership language. Various measures of social network position and language use confirmed this model indicating quantifiable metrics to infer online community leaders. Matthews et al. [39] developed Community Insights, a tool that supports effective leader behaviors by offering actionable analytics. That work also devised new community success metrics based on members' perceptions of how well the community supports their goals as well as members' overall satisfaction along with new visualizations that promoted community health. Such measures contrast with more standard quantitative behavioral measures such as levels of posting and community population growth.

2.2.2 Member Behavior. Many studies try to objectively contrast differing behaviors of leaders and members but results are often not definitive. For example, prior work using inferential methods finds that predicted leadership behaviors occur in members [46, 49, 62]. Other work describes role behaviors and transitions between roles. It is well-known that a small number of participants actively volunteer information while others consume it [44]. Preece and Shneiderman [49] argue that member participation follows a lifecycle apprenticeship model progressing through increasingly demanding roles: reading, contributing, collaborating, and leading. Again, we should note that most such lifecycle models are descriptive, with few studies actually operationalizing and measuring these stages. Furthermore, evidence for these apprenticeship transitions is mixed, for both definitional and empirical reasons. First distinctions between roles are not clearly agreed conceptually, making

it hard to determine exactly when such role shifts have occurred. And even allowing for these definitional concerns, empirical evidence suggests that participant behaviors are often stable [13]. Panciera et al. [45] examined communities of Wikipedia contributors, finding that a significant subset of 'newbie' participants showed typical leader behaviors in their initial contributions. These participants enacted leader-type work from their first day of usage without apprenticeship through a series of increasingly responsible stages. And work by Johnson et al [24] suggests that there is a subset of community members who show consistent leader-like behaviors across multiple contexts again arguing that role shifts may be the exception rather than the rule.

In this paper we provide quantitative data about differences between the social roles of leader and member with respect to content management by referencing. We examine role differences over time. From this, we can test an explicit assumption in most lifecycle models that members take increasing responsibility for referencing over the lifetime of the community with leaders' contributions becoming less important as the community matures. As prior work has shown, proactive behaviors by members are common in communities that are more successful, and theoretical models make the claim that members should conduct such behaviors as the community ages. This paper aims to test this claim. Technical characteristics of our data mean that we avoid some of the issues in defining and identifying role-specific behaviors that have been observed in prior studies.

2.3 Content Management and Hyperlinks

Various arguments have been made that collecting, organizing and actively maintaining content is critical to online communities [44, 48], with members being more likely to use communities that provide easily accessible information [38, 53, 58]. Well-organized content is also claimed to help retain members over the long-term [44, 48, 53] while disorganized content is argued to cause people to leave [17, 48, 53]. Tedjamulia et al. [53] argues that long-term participation depends on the community providing enough content, as well as the community's ability to leverage technology to provide access to that content.

Content management covers a range of different activities, including quoting already-existing content or summarizing prior useful content using FAQs [17, 20, 33]. In addition to these high-level management activities, there are also simple but pervasive methods such as referencing content through hyperlinks which are the focus of the current paper. Linking is commonly used for knowledge management across multiple tools including wikis, blogs, and forums [15, 56]. Hyperlinks were originally envisioned as a mechanism for both annotating individual documents and also indicating relationships between documents [43]. At the same time, links provide a straightforward way to navigate within and between document sets [8]. Much work has examined the uses and benefits of hyperlinks, showing that they serve to connect communities around similar content [15, 16, 56] and filter the abundance of content on the web [27, 28, 60].

Within wikis, hyperlinks are considered a fundamental aspect of content management as they connect topics and create context for those topics [4, 56, 57]. They also encourage cross-referencing, creating a navigable linked structure for networks of online resources, for example in educational contexts [15]. Within blogs and forums, hyperlinks are used as a resource for interlinking related ideas, typically associated with recommendation and summarization of said referenced content [8, 15, 28, 60]. While acknowledging the diversity of content management practices, in this paper, we use hyperlinks as a measurable indicator of active content management behaviors. We profile and analyze links based on different content sources to ensure we are measuring those used to actively manage content within and outside communities.

3 METHOD

3.1 Research Context

This research was conducted in a global technical enterprise offering technology products and services to businesses. The company widely encouraged employee leadership of, and participation in, internal online communities by making easy-to-use commercial community technology available to all employees. For the remainder of this paper, this enterprise community application will be referred to as "Communities". Communities is a pre-existing corporate product and was not developed for the purpose of the current study. A screen shot of the Communities landing page for a complex community working to launch a product is shown in Fig. 1. The Figure shows how that community synthesizes different resources concerning Web Marketing. Information is also shown about forum discussions, tags and community members, as well as bookmarks to related sites. All the communities we studied used the Communities application, which enabled participants to easily create a community space combining various social tools, e.g. forums, blogs, wikis, files, and bookmarks.

Our focus here is on content management practices involving referencing in these social tools. Across the thousands of participants and communities we sampled, skillsets varied widely as people were drawn from all across the company; skills ranged from highly technical (Software and Electrical Engineering) to less technical (Human Resources, Marketing and Management). Participants included both community owners and members, and we provide more details about these roles below. All participating users were employees of the same enterprise and were aware that this tool was being used for research purposes and proactively agreed in email to have their anonymized survey and logfile data used for analysis.

The software we studied was used for a very broad range of activities, from organizing social events in Communities of Practice, to goal focused activities like developing a marketing strategy for a range of new products across the company. The community activities we observed were very similar to those described in prior online communities research [9, 23, 49, 58]. Consistent with that other work, the communities we analyzed mainly focused on distributing knowledge, sharing expertise, answering questions and providing social support. Some communities appeared to be large communities of practice for members of a shared discipline (e.g. software engineers or marketers). Other communities appeared to be teams with executive leadership and more narrow goals specific to enterprise needs. Yet others were focused on specific recreational or technical problems over a shorter time-frame. To evaluate whether our communities functions overlapped with those identified in prior communities research, we surveyed community owners asking them to describe prevalent community activities and the type of the communities they managed. Owner responses were qualitatively clustered as follows: 41.1 % Communities of Practice (many members, mainly expertise sharing and networking), 29.4% Teams (executive leadership, fewer members, goal oriented projects), 3% Technical Support (providing technical advice to end-users), 1.4% Others. These subtypes both match those described in other analyses of enterprise communities [41], as well as the literature more broadly [23, 32, 58]. In what follows, we will not present systematic analyses of different community types as our analyses suggested few differences between types, although contextual analyses [41] suggest such differences may exist. The median number of members per community was 850, although as in prior research [9, 32], there was considerable variability (95% CI [765.08. 934.27]). Many employees were members of multiple communities. In summary then, the communities we studied involved varied participants and replicated many of the usage patterns that have been observed in other research on internet communities.

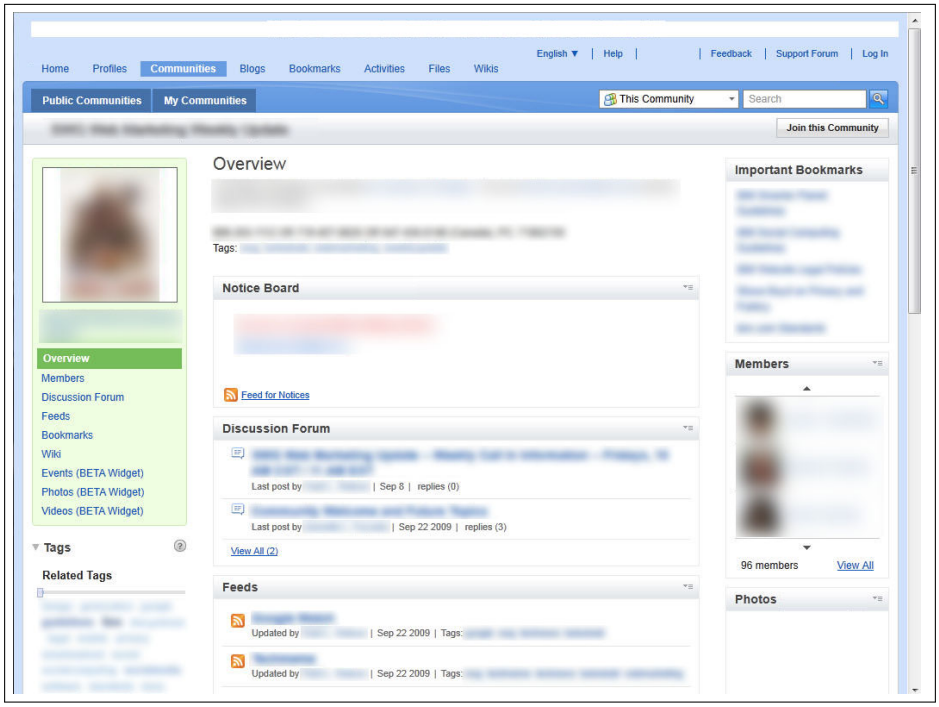

Fig. 1. Screenshot of a 'Communities' landing page, where participants gather content, discuss strategy and compile data in relation to coordinating Web marketing. Personally identifiable information has been blurred out. Overview of the community's media tools are provided in the left pane, recent discussions are within the center pane, and bookmarks, and community membership information are provided in the right pane.

3.2 Roles

We have seen that prior work proposes different roles for community contributors, but without clear consensus about how roles are behaviorally defined [18, 45, 46, 49]. Prior studies have also often employed inferential methods to distinguish leaders from members [24, 62]. However, we take a different approach as our data has unique properties that avoid some of these problems with inferential methods. In the Communities system, participants are officially designated to one of two roles: Owner or Member, each with different privileges. Members can view and post content with any tool, but may only edit their own content. They can also reference or link to others ' content. Owners are considered leaders as they have Members' rights but they can also edit any content, add/remove members, and configure tools. Owners are defined at community inception and do not change. These role definitions mean that our dataset has a gold standard for identifying social roles, removing the need for inferential methods and allowing direct measures to be made about role effects on specific behaviors of interest in relation to content management. Using a fixed definition of roles, we can examine role-specific Owner and Member management behaviors over time. While a fixed definition of roles may suggest limitations, prior empirical work shows peoples' online roles tend to be relatively static [13, 45]. Although our Members lack certain privileges, in common with internet-based communities there is still considerable room for them to display important leader-like content management behaviors. For example, they can create links to manage forum, wiki or blog content. Of direct interest in this paper is the extent to which these individuals within

the fixed Member role begin to take responsibility and show such content referencing behavior as the community matures and content begins to accumulate.

To check for consistency between our system-designated Owner and Member and definitions of roles used in prior work, we compared typical role-specific behaviors identified in those prior studies, relating to networks, language use and identity behaviors. Consistent with that prior work, Owners had larger communication networks [24], higher usage of social linguistic styles and adopted the community identity [24, 62], when compared with regular Members.

In this paper, we therefore test theoretical predictions about increasing Member responsibility for content referencing [23, 49]. We assess this by observing whether Members' target content management behaviors i.e. linking, increase with community age as hypothesized by lifecycle models.

3.3 Community Sampling and Data Collection

Our criteria for including enterprise communities in our analysis were:

- *Active management*: Leaders had to sign up for Community Insights [39], a tool to help leaders enhance their community. A research goal is to make community design recommendations, so we wanted active leaders.
- *Active posting*: Updated in the last month. We wanted successful extant communities since our aim was to describe effective usage practices. We did not include communities that were inactive.

There were 2,010 communities that met our criteria of being active, generating a total of 428,476 posts and 1,246,570 links. Recall that the Communities system was originally developed the purpose of providing a platform to establish healthy enterprise communities that support employees and corporate processes. We were therefore able to collect log-files data on every user interaction, pages viewed, clicks on interactive widgets, from July 2007 to May 2014. This data was then linked to demographic Communities data and logged in a MySQL database. For each post, with participants' agreement we captured:

- Community ID (Where it was posted)
- Author ID (Anonymous Unique identifier)
- Date (Time stamp when post was made)
- Tool (Which tool the post was in: e.g., blog, wiki, etc)
- Role (Member vs Owner of community posted in)
- Date of Community Creation (To determine when in the community lifecycle the post was made)

Our focus was on links for content management, and for each link we captured:

- Source Community (Where it was posted)
- Source Tool (Which tool the link was posted in)
- Target Location (Internal or external to source community)
- Target Tool (The tool the link points to)
- Date (Time stamp when link was posted)
- Author ID (Anonymous Unique identifier)
- Author Role (Member or owner)

3.4 Measures

3.4.1 Creation and Referencing. Creation was defined as new content added to a community. Content can be added in different ways using different tools. Tools available to the community

were: forums, blogs, wikis, and bookmarks. Our measure of content created was the sum of the number of forum posts and replies, blog posts and replies, wiki edits, and bookmarks.

We define referencing as the act of linking to already-created content and potentially annotating it for other community members through the use of a hyperlink in any of the social tools. From prior work, we know that hyperlinks are used to reference external information sources that are relevant to community discussions and to organize content within the community [14, 38, 42, 54]. As noted earlier, referencing has been successfully used in prior work to assess content management, but it is not an exhaustive measure of content management, as it excludes behaviors such as quoting or summarizing prior content using methods such as FAQs. Nevertheless, we chose to measure content management via referencing, as it can be more reliably operationalized as an indicator of content management than those behaviors; hyperlinks are easily countable and extractable from posts. To validate links as a measure of content management, we also report a qualitative analysis showing that specific link types are reliably used for referencing.

Hyperlinks in our data have different sources they are referencing. We are interested in examining references that link to existing content within the community or another community. We identified these hyperlinks labeling them as Internal (hyperlinks referencing content within the community) or External (hyperlinks referencing content from another community but still within the enterprise intranet Communities app). Overall, we found 205,693 (16.5% of total hyperlinks) Internal references and 313,922 (25.1%) External references. Additional types of hyperlinks that exist were Enterprise intranet but outside of the Communities app (45.2%) and those referencing the Open Internet (12.8%). The remaining 0.4% of hyperlinks were unidentifiable. Below we present two analyses of link functions. We first conducted a machine learning analysis showing that links to the 4 different sources (Internal, External, Enterprise Intranet and Open Internet) involve distinctly different content. A second qualitative analysis examined the content management functions of these four different link sources, showing that only Internal and External links are actively invoked for content management.

3.4.2 Temporal Analysis and Lifecycle. Our focus is on whether and how community behaviors change over time. Prior work has proposed different community lifecycle phases. However, these phases are extremely difficult to operationalize, e.g., how might we determine that a community has moved from inception to growth or from growth to maturity [23]? Communities may also develop at different rates, making it difficult to compare between them. Rather than proposing ad hoc behavioral indices for these phases, we collected long-term data over communities for a 36 month period in relation to their age. Not all communities analyzed were 36 months old but were included in the aggregated behaviors up to their age when the data was collected, for example a community that is only 15 months old at the end of our data collection would be included in the data for months 1-15 but not for months 16-36. The average community age was 24.2 months (95% CI [23.42, 24.98]).

We analyze time relative to the creation date of each community, for example, month 1 indicates the behaviors of all communities from their creation to age 1 month. This minimizes the possibility of outside events influencing aggregated behaviors across multiple communities. Outliers at each time step were filtered using a Median Absolute Deviation [35]. For each behavior, we first examine general trends by fitting a local polynomial regression to the time series of all communities. Polynomial methods are used because linear models offered poor fits and provide a stronger visualization of changes over time, although they don't allow for statistical comparisons. We use a nonparametric regression since these time series analyses were found to be non-normal. 95% confidence intervals are plotted along the regression lines. To compare roles across time, for each target behavior we separate those that were conducted by Owners and Members for each month

and fit another local poly regression for each role to visualize the differences. We then used a mixed model regression to evaluate statistical significance between roles.

Some of the behaviors we analyze are relatively infrequent, occurring a few times per month. However, it is important to note that all the communities analyzed are still active when the data was collected suggesting that even low levels of behavior are markers of long-term community survival.

4 RESULTS

Before assessing our main research questions, we first present an analysis of link usage to assess whether links reliably assess important aspects of content management. We begin with a machine learning analysis of different link sources, showing that sources are distinct, followed by a qualitative analysis indicating that only Internal and External links are being used for content management. We then explore both posting and referencing over time. We then evaluate role changes by comparing leaders and members in their posting and referencing behaviors.

4.1 Referencing Link Sources For Content Management

4.1.1 Link Sources are Distinguishable by Machine Learning Classification. Recall that there are four sources of links based on whether they reference content that is Internal to the community, External, outside the Communities app but within the Enterprise Intranet, or from the Open Internet. We first conducted a machine learning experiment on these 4 link sources to determine if links differ depending on the source content they reference, by evaluating the words used around each link. We extracted the content from the sentence before the posted link, as well as text that contained an embedded link. We then extracted a series of N-gram (1-2-3 gram) features from the text. We fitted a Support Vector Machine [40] to this data and used a 72-18-10 data split for training, validation, and test sets for modeling training. The validation set was generated using a 5-fold Stratified cross validation procedure. Stratified cross validation keeps the distribution of classes equal through the data splits, helping address unrepresentative data splits [30, 61]. To evaluate the model, we used Precision (True Positives divided by True Positives plus False Positives), Recall (True Positives divided by True Positives plus False Negatives), and F1-Score (a combination of precision and recall) on each link class [47]. Link class performance for the model is ranked based on F1-Score.

Table 1 shows the results. The model performed well with an overall F1-Score of 0.68, indicating a reliable difference in the type of words used in the sentence before the link, showing that referencing behaviors are reliably different based on the source of the referenced content. The model performed best in identifying link types in the Enterprise Intranet class (highest F1-Score), and Open Internet links were also classified well. Both Internal and External were fairly accurate in their classifications as both are still above baseline (F1-Score > 0.25). Furthermore, confusion matrix analysis indicated that the majority of misclassifications involved Internal and External classes being confused for each other suggesting their functions overlapped. We return to this overlap in our qualitative analysis.

Our machine learning analysis suggests that the 4 link source types are distinct, so we then went on to qualitatively analyze the functions of each link source to identify whether and how they were used for content management. We explored example posts of each link source category, namely Internal, External, Enterprise Intranet and Open Internet links. Example posts show how different sources of referencing links managed content. Explicit URLs are indicated by "<HYPERLINK>" and embedded references by hyperlink tags bookending the text "<HYPERLINK> </HYPERLINK>". Personally identifiable information has been anonymized.

Table 1. Results for Classification of Link Types using a SVM model on N-gram feature set, showing that link sources reliably index different types of content. Enterprise Intranet has the best performance based on F1-Score, with Open Internet having the second best performance. Internal and External had the lowest performance but were still better than random.

Link Source Classification			
	Precision	**Recall**	**F1-Score**
Internal	0.63	0.32	0.42
External	0.66	0.59	0.63
Enterprise Intranet	0.69	0.85	0.76
Open Internet	0.75	0.61	0.67

4.1.2 Internal Links To Reintroduce Existing Content. Internal references were commonly used by experienced users to draw attention to prior material within the community that relates to a new post. Referencing was done for the benefit of newer users who seemed to be unaware of that existing content [32].

> Hi User 1, that 's a very good idea. We actually have a "XX" forum where everyone can access and trade their stuff. You should check it out too. Here 's the link <HYPERLINK>

This post links to existing community content that the newcomer User 1 seems unaware of. The explanatory text labels the prior content via a short description, and the reference provides direct access to that content. Using the link serves two important content management functions. It avoids duplicating prior content and so reduces the accumulation of content within the community. Use of the link also explicitly signals relations between content in different parts of the community, in this case between the current post and the "XX" forum resource. Other Internal references were similarly used to announce new content to the community, where that content is being posted in a pre-existing community resource.

> User 2 has announced that User 3 is the new crucial position... Read the announcement on the < HYPER-LINK>YY Community Wiki</HYPERLINK> here: < HYPERLINK>

This post again serves multiple referencing functions. As in the prior example, the embedded reference to the 'YY Community Wiki ', reminds readers about the existence of that local community wiki resource. The link also provides ready access to the content of that announcement. Again using the link reduces content accumulation, as content is not duplicated within the current post, but can be accessed from the Wiki by those interested in the announcement details. Other posts used internal references to promote group action, in this case a forum for community brainstorming.

> Foster the collaboration. If you have ideas on Tools saving, pls. put them here -> < HYPERLINK >. If you have any further questions please let me know.

This post is proactive in encouraging new community posts as opposed to organizing existing information. Nevertheless, it uses the same content management approach as the prior examples; it directs the reader to existing internal community resources where they should post new content without redundantly duplicating a detailed prior description of those resources. Other referencing behaviors promoted content outside the community. We now characterize the functions of these different types of outside links.

4.1.3 Functions of External Links. Many External links overlap with functions already identified for Internal links. External links identify directly relevant resources in other enterprise communities that help the local community better organize their own information.

> Some more info i have [sic] digged up is a wiki with guidance: It is from the <HYPERLINK>Community Builders Wiki</HYPERLINK>: Community Leader tips from social science research. That page describes what you should to do in advance and after creating a community, to ensure it becomes successful. A lot of questions I have seen (like small or big) are answered by that material indirectly.

Here the poster promotes external information they believe is relevant to their own community. They briefly summarize the content of the reference, justifying why it is relevant. This use of referencing means that content is not duplicated or proliferated within the community. Other posts involving external references aim to organize material relevant to the local community, with less focus on explaining the referenced external content.

> Welcome to the BC Tools Focal Point Topic in the BC Focal Point Forum Some tools links: <HYPERLINK> <HYPERLINK> <HYPERLINK>

This post sets up a simple structure to reference external content while directly imposing an active organization on that material. These examples of internal and external links indicate how referencing is used to achieve important aspects of content management. Each case involves a combination of linking to outside community content, while simultaneously summarizing or annotating the content. Use of the reference link avoids content duplication within the community reducing accumulation of content over the long-term. Most importantly links organize or impose structure on that internal or external content.

4.1.4 Enterprise Intranet and Open Internet Links. Other forms of linking outside the community have different functions that less directly involve content management. Enterprise Intranet links reference the corporate intranet and can identify important resources such as programs or events. These resources can assist with tasks users are trying to execute, but tend not to actively relate to community content:

> Hi User 4, I suggest two panels being opened when viewing a document in this tool 1) All open docs for tool goes like this (example): <HYPERLINK> 2) Doc Journal which goes like this (example): <HYPERLINK> Obviously you 'd need to replace the value, second value or number with the one at hand

A final class of links reference the Open Internet. Open Internet links do not point to existing organizational resources. Instead they usually identify additional external information, for example identifying the person introduced in the post with a link to their personal webpage. These Open Internet links usually do not reference organized existing information or provide follow-up actions for readers. Internal and External links reference community content, providing active resources for organizing and sharing target community information. In contrast Intranet and Open Internet do not seem to actively address community organization the way that internal and external links do. We therefore limit our quantitative analysis to Internal and External links, i.e. posts that link within a community or to another related enterprise community.

4.2 Overall Lifecycle Trends in Creation and Referencing: Total Content Increases but Linking Decreases.

Our qualitative analysis characterized referencing for Internal and External community links showing active content management. We now examine these specific management behaviors over time, using quantitative methods. In each of the following analyses, we compute a time series trend by fitting a local polynomial regression for each month. We report analyses of absolute rates of posting and linking, but similar analyses that normalize these rates by community show similar results. We have also conducted different analyses exploring whether there are differences

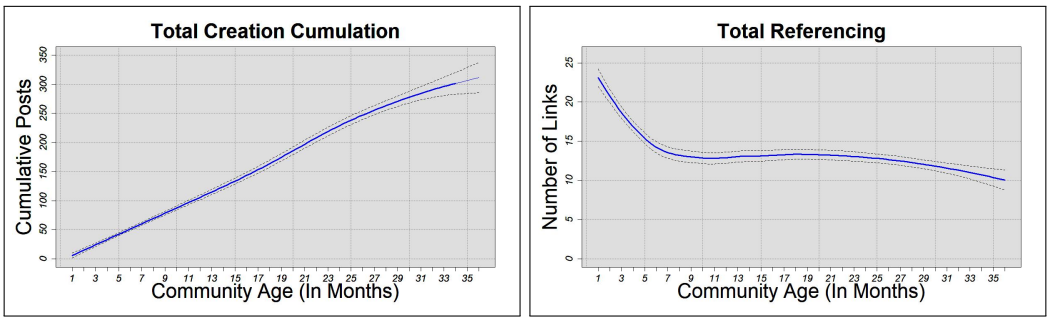

Fig. 2. Communities accumulate ever-larger amounts of content over time. Left Fig. shows a Local Poly Fitted Regression of cumulative posts/month (95% confidence intervals shown as upper and lower dotted lines). Right Fig. shows a local poly fit of referencing within communities over community age: (95% confidence intervals shown as upper and lower dotted lines) Referencing starts high but decreases until around 15 months, then remains steady.

in creation and referencing for different types of communities and whether there are differences between different tools (blogs, wikis, forums, bookmarks). There were no noteworthy differences between community types, so in the interests of space we do not report these results. There were differences between tools. As we anticipated, wikis were used more often for referencing, which we discuss later. Following claims in prior work [23, 49, 58] we examine whether active referencing activities are a response to accumulating content. We therefore begin by characterizing content creation, i.e. how content changes over time.

4.2.1 Content Accumulates Over Time. The left plot in Fig. 2 shows a local polynomial regression on the cumulative content for all communities by month. Over time, communities acquire ever-larger amounts of content. Overall, content shows a steady visual linear accumulation with a slight increase as communities approach 20 months, dropping slightly in months 32 and 33. This content accumulation would seem to demand greater organization; communities on average have about 3.5 times the amount of content at month 36, compared with month 10. As communities mature, we should therefore expect them to engage more actively referencing activities to manage this accumulated content.

We evaluated the relationship between total content and active organizational linking using a repeated measures regression with communities as subject. Somewhat unexpectedly, and contradicting lifecycle models, this analysis showed a very weak relationship between cumulative content and linking as the beta weight is extremely low, (β = 0.008, SE = 0.001). Even though this relationship is statistically significant ($p < 0.001$), the small coefficient indicates that cumulated content has little predictive power in explaining linking behavior. We therefore went on to examine linking behavior directly to explore why this is the case.

4.2.2 Referencing Rates Drop Over Time. To assess referencing visually, we plotted a local poly regression of number of links/month over the age of the community in the right plot of Figure 2. To our surprise, linking rates decreased over time. Despite each community having many more posts to organize, linking rates dropped over time. The reason for this drop becomes clear when we examine trends in referencing. Most referencing occurs within the first month. We see a large negative rate of change in the first few months but this approaches 0 around the end of the first year. Consistent with prior qualitative work on communities [39], Figure 2 suggests an intense referencing phase in the initial months. This start-up phase involves active referencing as participants seek to proactively

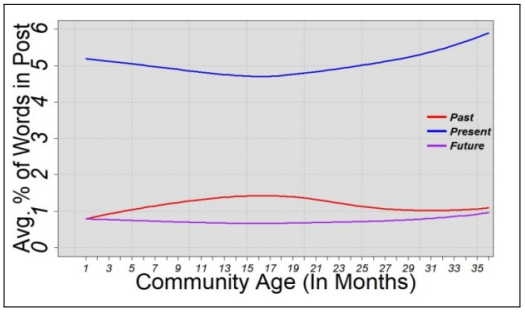

Fig. 3. Results of analysis on link temporal focus. Left shows the results of examining language tense in posts containing at least one link. Posts contain more Present than Past and Future tense words.

organize early content to allow other members to effectively navigate that content. After the startup phase, we expected the growing weight of accumulated content would demand active organization, leading to an increase in referencing over time. Instead we found that referencing decreased after the first month, levelling off after 10 months, despite ever-accumulating amounts of content late in the community's lifespan. Our results over 36 months therefore challenge one aspect of those models which assume that content management increases over the lifetime of communities. A possible explanation for this might be that communities tend to focus on content that is more recent and ignore older accumulated content thereby removing the need to organize that overlooked older content.

4.2.3 Referencing Has a Recency Bias. To test that possibility, we examined posts that contained at least one link and explored if these posts are more focused on present rather than past content. Examining the content of a post that contains a link, we used the tense categories from the tool Linguistic Inquiry and Word Count (LIWC)[52] as a proxy for focus within the post and examine their use over time. Fig. 3 shows the average Past, Present, and Future language use over community age reported as a percentage of words in a post. Posts were far more likely to focus on Present than both Past and Future tenses. More significantly we saw no increases in Past tense usage. If participants were actively referencing past archival content we might expect Past tense references to increase throughout the community lifespan as past content accumulates. Overall this is consistent with a recency based content focus. While this tense analysis shows that link posts are more focused on the present, it doesn't directly examine the age of older content that is being referenced. To assess this, we looked at instances of where links were repeated. We measured the median time difference between repetitions of the same link. In the last 6 months of the community lifecycle, the largest median time difference between repetitions was 27 days, a 2% time window within a community lifecycle of 36 months, indicating that communities are focused on recent content. Overall, tense and link repetition analyses argue that referencing content has a recency bias.

4.2.4 Role Differences: Members are increasingly responsible for content creation but not for referencing. Our second research objective was to examine Lifecycle models which claim that community responsibilities shift over time. Those models argue that, compared with community Owners, Members take an increasingly active role over time both creating and managing content. We therefore expected members to engage in higher rates of linking as the community matured. To establish baselines we first analyze content creation rates in members versus owners. Recall too that Members and Owners are formally defined in the communities that we are analyzing.

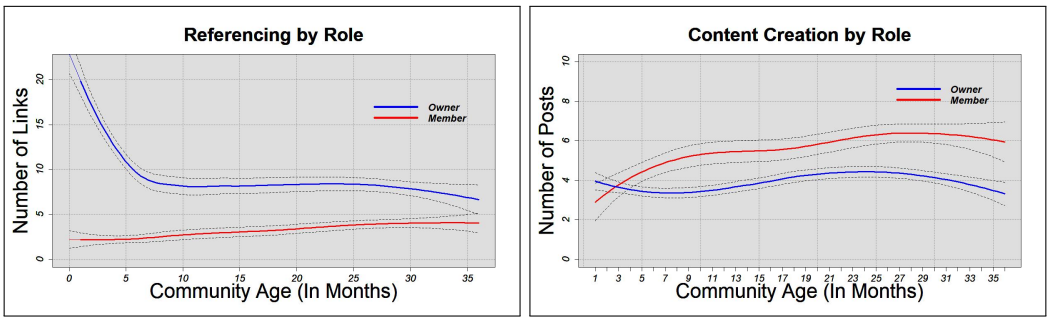

Fig. 4. Members dominate content creation after first few weeks. Right Fig. shows local poly fit of creation by roles. Left Fig. shows local Poly fit for role referencing over time. Owner dominate referencing throughout, but quickly decrease after a few months and then have a gradual decline. Members referencing remains relatively similar throughout.

4.2.5 Members Dominate and Take Increasing Responsibility for Content Creation over Time. We compared the creation behaviors of Members versus Owners (see Figure 4). The Figure indicates that Owners show higher content creation rates only at the community's inception. As expected, the average number of Member posts increases over time, whereas Owner posting rates remain steady. This is consistent with theoretical lifecycle accounts arguing that members take increasing responsibility for creating new content as the community matures. From early on, Members drive content production, increasing production faster than Owners, but both Members and Owners decrease production rates as the community matures after 30 months.

To test the significance of this difference, we used a zero-inflated negative binomial mixed-model regression. Poisson and Binomial regressions are typically used for fitting count data and due to the unequal values of the mean and variance for each time point in the data, binomial distributions produced a better fit ($p < 0.0001$) and were not over dispersed. We used a zero-inflated model, as roughly 55% of the data at each timestamp are zeros. This type of model accounts for high proportions of zeros by considering them in a separate process. A mixed model was used to account for individual community variance. The libraries lme4 and glmmADMB in R were used for modeling [1, 6, 7], the zero-generation process was handled by the glmmADMB library which treated the zero-count outcome as a mixture of structural and sampling zeros. The model applies this as follows: given the outcome of the model (Y) and the probability of the outcome being equal to zero being p, a proportion of Y of size p comes from extra zeros and a proportion of Y of size 1-p comes from the Binomial distribution. The density of this model can therefore be characterized following Böhning et al. [10] as:

$$f(y : p, \mu) = pBN(y, 0) + (1 - p)BN(y, \mu) \quad (1)$$

The equation 1 shows the model is a mixture of two classes, the first class having a fixed value of 0 and the second being the Binomial distribution (indicated by BN in Function 1) [1, 10]. Our p for this model would be 0.55, given our sample. Such a model has a high computational demand, we restricted the model to run on a random sample of 200 communities. The results of the models are in shown in Table 2. The model shows differences in both Role and Time, with an interaction between the two. This is consistent with the right plot in Figure 4, showing overall differences between Roles with members creating more content and Time as overall differences in content created increased by month. The Role by Time interaction follows from Members overtaking Owner's initial posting rates after month 2.

Table 2. Results for Mixed Model for Community Roles For Creation

ZF Negative Binomial Mixed Model Regression Creation Results			
	Estimate	**Std. Error**	**P-value**
Role	0.578	0.107	6.4e-8***
Time	0.044	0.011	8.2e-5***
Role * Time	-0.015	0.006	0.012*

Table 3. Results for Mixed Model for Community Roles For Referencing.

ZF Negative Binomial Mixed Model Regression Referencing Results			
	Estimate	**Std. Error**	**P-value**
Role	1.751	0.121	< 2e-16***
Time	-0.026	0.013	0.047*
Role * Time	-0.044	0.007	5e-10***

4.2.6 Members Don't Increase Referencing Over Time. We next analyzed referencing behaviors comparing different roles. Following lifecycle models, we anticipated that Members' referencing would increase as they assumed greater responsibility for content management compared with Owners. However, this was not confirmed. In contrast to creation, where Members increasingly dominate, the right plot in Figure 4 shows Owners show greater referencing and this is maintained throughout. Using the same modeling procedure as before, Table 3 again shows main effects of Role and Time and an interaction between Role and Time. Consulting the right plot in Figure 4, we see that differences between Roles arose because Owners created more links than members. Effects of Time arise because overall linking decreased over time, again contradicting lifecycle models. The Role by Time interaction results from the decrease in Owner linking after the first few months, as Owners shift from intensive early link creation, whereas Member linking is relatively stable throughout the community lifespan.

4.2.7 Analysis of Wiki Usage. To confirm our referencing results, we conducted a second exploratory tool-centric analysis. Recall that we collected data from several social media tools. We noted earlier that there are multiple methods to support content management, with wikis being a tool that is commonly used for this purpose [38, 56]. To check and extend our link analysis, we therefore also assessed whether wiki tool usage was consistent with the link referencing behavior we had observed across all tools. We found that the rate of wiki creation remained constant across the community lifespan and did not increase over time. There were no differences in the mean number of wikis created when comparing between the first and second 18 months of the communities' lifespans, using a Kolgormorov-Smirnov test ($D = 0.3129$, $p = 0.2284$). Turning to role behaviors for wikis, we found Owners on average created more wikis than Members, with Owners creating 10.8 wikis and Members creating 0.3 wikis. With respect to time, we again found that Owners consistently created more wikis than Members throughout the community lifespan ($D = 0.8649$, $p < 0.0001$). Overall, then, this analysis of wikis confirms our link data. Similar to the results for overall linking, we see little increase in use of wikis overall; Owners dominate wiki referencing with no significant changes in Member wiki creation over time.

4.2.8 Limitations. There are limitations to this initial exploratory analysis. As we noted, we restricted quantitative analyses to one operationalizable aspect of content management using

hyperlinks. Although we present other consistent exploratory data from wiki usage, it could be that using these link measures leads us to underestimate other content management behaviors and tool usage. Furthermore, communities may turn to external tools to manage content that we were unable to measure. We also rely on simple binary distinctions between Owners and Members and we are aware that theoretical work proposes more subtle differences in community roles [23, 49, 58, 62]. However, these decisions were motivated by the need to gather reliable quantitative data to make comparisons, acknowledging the difficulty of accurately operationalizing prior qualitative definitions of complex community behavior. Furthermore, our analysis is limited to the first 36 months of interactions, and it could be that referencing only becomes critical later in a community's lifespan, although lifecycle models would have to be modified to incorporate this claim. Finally, this work explores referencing in enterprise settings. Although past work shows strong overlaps in community behaviors across contexts, future work should explore whether our results generalize to internet communities.

5 DISCUSSION

While this study is exploratory, our findings are nevertheless counterintuitive. First, we did not find the expected increases in referencing over time predicted by influential community lifecycle models [23, 49, 58]. One possible explanation is that after an early burst when a community is initiated, referencing ignores older 'stale' content and instead focuses on recent active content rather than trying to organize the entire set of community content. Data from tense usage and link repetitions are consistent with this recency bias. An alternative explanation could be that older references are handled through other content management methods, e.g. external storage systems (version control, file storage or synchronization services like Google Drive). However it seems unlikely that communities would link to external Wikis when they have the more straightforward alternative of using their own community wiki for content management. We were able to test alternate content management methods to linking that involved community wikis, and our exploratory results confirmed our referencing analysis. Content management using external methods (e.g. versioned repositories, external wikis), could of course have occurred and be the cause for this drop in referencing. Future work is needed to examine more specific open internet referencing using external management tools.

A second research question concerned division of labor between Members and Owner roles and how this changes over time. Consistent with lifecycle models, Members were increasingly responsible for content creation over time. But contrary to those models, referencing was managed by Owners throughout the life of the community. This discrepancy in management is notable given that there were almost 100 times as many community Members as Owners, and that content has increased many times over during community's lifetime.

Another important question is the extent to which our results generalize to other online community contexts. We have presented data on community types, leaders and behaviors which seem to indicate that our work is representative of other well-studied online communities. First our data for content referencing replicate known power law effects with the majority of referencing effort being contributed by relatively few [45, 59]. Second our data includes many commonly occurring community types observed elsewhere, such as communities of practice, teams and small work groups [41]. Furthermore, although our data used fixed as opposed to flexible participant roles, an analysis of Owner behaviors in our sample revealed strong overlaps with prior reported norms for leader behaviors [24]. Finally our analysis of hyperlinks matched use cases detailed in prior work [14, 56]. Overall these observations suggest clear consistencies with other online communities' research, giving us confidence that results will generalize elsewhere.

Our results on reduced referencing and recency bias are important for theory and practice. These phenomena may not have been observed before because there have been relatively few long-term analyses of long-term community behaviors and little focus on referencing. Theories need to incorporate our results, explaining why communities apparently fail to organize and refer to accumulated prior content. It could be that communities are inherently biased to focus on present discussions rather than extensive past content. This bias may reflect short-term participation with members joining the community but leaving after relatively short periods of active contribution. Such 'churn' would make it hard for communities to build a shared long-term perspective on their content. It may also be that members find it hard to use tools such as wikis that promote referencing and we return to this point below.

Our other results extend prior literature on distribution of labor within communities. Referencing follows the well-known power law distribution [45, 59] where only a select few users (Owners) conduct the majority of work. However, this was not the case for content creation as the much larger group of Members enacted the majority of work. It may be that referencing follows a pattern similar to Panciera et al. [45], where Owners' initial commitment to the community leads them to be relatively more active than Members throughout the community lifespan. However it is also clear that Owners' levels of referencing drop as the community ages, which may reflect burnout, or an unwillingness to organize content created by others. Again these are interesting questions for future research.

Our results have implications for theory as well. First, they suggest a need to refine lifecycle models to include a greater emphasis on actual community practices, in particular to incorporate our findings about decreased referencing over time. Second, Owners were mainly responsible for content referencing over the life of the community; this suggests the need to educate and encourage Members not only to create content but also to reference content. Third, while Owners' referencing is consistent with standard power law accounts [59], this wasn't the case for creation where Members posted more content than owners. This suggests a need to refine power law accounts to include the specific tasks involved in the community. Overall, our findings on temporal characteristics generate new questions about lifecycle theories, suggesting the design of new tools and metrics to assess community success.

Finally, these results inform the design of new tools and metrics for community building. In particular, organizational tools such as wikis and bookmarks might be designed to encourage more active participation by members. Prior work shows that content management activities play an important usability role for community members [38] and that content management tools, i.e. wikis and bookmarks, are currently challenging for members to deploy [44]. Automatic text processing methods could also assist in referencing, for example by summarizing existing content. It may also turn out that community's recency bias means that content management tools only need to focus on newer information. An alternative design approach may be to modify community members' recency bias by designing new interfaces that draw their attention to interesting older content. One solution might be to model the approach taken by Facebook's "On this Day" [3] or Google Photos "Rediscover this day" [2], which both focus on re-presenting older content relevant to recent activity or content that previously received extensive active feedback. For example, communities might resurface older posts based on their direct relevance to recent posts, or more simply because the re-presented posts promoted highly involved discussion in the past. Of course such re-presentations would need to be carefully designed so that users are aware of the motivations for resurfacing older posts. Another design approach might encourage community leaders to flag interesting content for later resurfacing. More automated solutions could involve detecting overlap between a new contribution and a successful prior contribution, leading the application to resurface the prior

solution. This is similar to work done on automatic answer detection in question oriented forums [12].

6 CONCLUSION

This paper presents new results about how community behaviors change over time. Although this study is exploratory in assessing content referencing, our findings are nevertheless counterintuitive in light of current theory. As expected, content increased over time and Members took responsibility for creating such content. However contradicting current life-cycle model predictions, referencing did not increase as content accrued and instead decreased as communities aged. Furthermore, Members did not assume responsibility for content referencing. We explored one possible explanation for this lack of content referencing, finding that communities tend to focus on very recent content, making it less important for them to focus on managing older material. We describe new tools and leader practices that might better utilize older content.

REFERENCES

[1] [n. d.]. The glmmADMB package. http://glmmadmb.r-forge.r-project.org/

[2] [n. d.]. Google Photos - All your photos organized and easy to find. https://www.google.com/photos/about/

[3] [n. d.]. Introducing On This Day: A New Way to Look Back at Photos and Memories on Facebook | Facebook Newsroom. https://newsroom.fb.com/news/2015/03/introducing-on-this-day-a-new-way-to-look-back-at-photos-and-memories-on-facebook/

[4] Sisay Fissaha Adafre and Maarten de Rijke. 2005. Discovering missing links in Wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*. ACM, 90–97.

[5] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn RosÃĭ, and Xiaoqing Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 959–968.

[6] Douglas M. Bates. 2010. *lme4: Mixed-effects modeling with R*. Berlin.

[7] Ben Bolker, Hans Skaug, Arni Magnusson, and Anders Nielsen. 2012. Getting started with the glmmADMB package. *Available at glmmadmb. r-forge. r-project. org/glmmADMB. pdf* (2012).

[8] Leanne Bowler, Daqing He, and Wan Yin Hong. 2011. Who is referring teens to health information on the web?: hyperlinks between blogs and health web sites for teens. In *Proceedings of the 2011 iConference*. ACM, 238–243.

[9] Brian Butler, Lee Sproull, Sara Kiesler, and Robert Kraut. 2002. Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work* 1 (2002), 171–194.

[10] Dankmar BÃűhning, Ekkehart Dietz, Peter Schlattmann, Lisette Mendonca, and Ursula Kirchner. 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162, 2 (1999), 195–209.

[11] Xu Cheng, Cameron Dale, and Jiangchuan Liu. 2008. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*. IEEE, 229–238.

[12] Gao Cong, Long Wang, Chin-Yew Lin, Young-In Song, and Yueheng Sun. 2008. Finding question-answer pairs from online forums. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 467–474.

[13] Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 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*. International World Wide Web Conferences Steering Committee, 307–318. http://dl.acm.org/citation.cfm?id=2488416

[14] J. De Maeyer. 2013. Towards a hyperlinked society: A critical review of link studies. *New Media & Society* 15, 5 (Aug. 2013), 737–751. https://doi.org/10.1177/1461444812462851

[15] Peter D. Duffy and Axel Bruns. 2006. The use of blogs, wikis and RSS in education: A conversation of possibilities. *Proceedings of the Online Learning and Teaching Conference 2006,* (2006).

[16] Henry Farrell and Daniel W. Drezner. 2008. The power and politics of blogs. *Public choice* 134, 1-2 (2008), 15–30.

[17] Robert Farrell. 2002. Summarizing electronic discourse. *Intelligent Systems in Accounting, Finance & Management* 11, 1 (2002), 23–38.

[18] Eric Gleave, Howard T. Welser, Thomas M. Lento, and Michael A. Smith. 2009. A conceptual and operational definition of'social role'in online community. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE, 1–11. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4755505

[19] Catherine Hall and Michael Zarro. 2012. Social curation on the website Pinterest.com. *Proceedings of the American Society for Information Science and Technology* 49, 1 (Jan. 2012), 1–9. https://doi.org/10.1002/meet.14504901189
[20] Derek L. Hansen. 2006. Knowledge sharing, maintenance, and use in online support communities. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1751–1754.
[21] Itai Himelboim. 2010. The international network structure of news media: An analysis of hyperlinks usage in news web sites. *Journal of Broadcasting & Electronic Media* 54, 3 (2010), 373–390. http://www.tandfonline.com/doi/abs/10.1080/08838151.2010.499050
[22] Paul Holmes and Andrew M. Cox. 2011. 'Every group carries the flavour of the admins': leadership on Flickr. *International Journal of Web Based Communities* 7, 3 (2011), 376–391. http://www.inderscienceonline.com/doi/abs/10.1504/IJWBC.2011.041205
[23] Alicia Iriberri and Gondy Leroy. 2009. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)* 41, 2 (2009), 11.
[24] Steven L. Johnson, Hani Safadi, and Samer Faraj. 2015. The emergence of online community leadership. *Information Systems Research* 26, 1 (2015), 165–187. http://pubsonline.informs.org/doi/abs/10.1287/isre.2014.0562
[25] Bob Kanefsky, Nadine G. Barlow, and Virginia C. Gulick. 2001. Can distributed volunteers accomplish massive data analysis tasks. *Lunar and Planetary Science* 1 (2001).
[26] Amy Jo Kim. 2000. *Community building on the web: Secret strategies for successful online communities*. Addison-Wesley Longman Publishing Co., Inc.
[27] Sheila Kinsella, Mengjiao Wang, John G. Breslin, and Conor Hayes. 2011. Improving categorisation in social media using hyperlinks to structured data sources. In *The Semanic Web: Research and Applications*. Springer, 390–404. http://link.springer.com/chapter/10.1007/978-3-642-21064-8_27
[28] Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 5 (1999), 604–632.
[29] Joon Koh, Young-Gul Kim, Brian Butler, and Gee-Woo Bock. 2007. Encouraging participation in virtual communities. *Commun. ACM* 50, 2 (2007), 68–73.
[30] Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Stanford, CA, 1137–1145. https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf
[31] Robert E. Kraut and Andrew T. Fiore. 2014. The role of founders in building online groups. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 722–732. http://dl.acm.org/citation.cfm?id=2531648
[32] Robert E. Kraut and Paul Resnick. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.
[33] Yu-Sheng Lai, Kuao-Ann Fung, and Chung-Hsien Wu. 2002. Faq mining via list detection. In *proceedings of the 2002 conference on multilingual summarization and question answering-Volume 19*. Association for Computational Linguistics, 1–7.
[34] Jean Lave and Etienne Wenger. 2002. Legitimate peripheral participation in communities of practice. *Supporting lifelong learning* 1 (2002), 111–126.
[35] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49, 4 (July 2013), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013
[36] Pamela J. Ludford, Dan Cosley, Dan Frankowski, and Loren Terveen. 2004. Think different: increasing online community participation using uniqueness and group dissimilarity. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 631–638.
[37] Tara Matthews, Jilin Chen, Steve Whittaker, Aditya Pal, Haiyi Zhu, Hernan Badenes, and Barton Smith. 2014. Goals and perceived success of online enterprise communities: what is important to leaders & members?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 291–300. http://dl.acm.org/citation.cfm?id=2557201
[38] Tara Matthews, Steve Whittaker, Hernan Badenes, and Barton Smith. 2014. Beyond end user content to collaborative knowledge mapping: Interrelations among community social tools. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 900–910. http://dl.acm.org/citation.cfm?id=2531694
[39] Tara Matthews, Steve Whittaker, Hernan Badenes, Barton A. Smith, Michael Muller, Kate Ehrlich, Michelle X. Zhou, and Tessa Lau. 2013. Community insights: helping community leaders enhance the value of enterprise online communities. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 513–522.
[40] Tony Mullen and Nigel Collier. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 412–418.
[41] Michael Muller, Kate Ehrlich, Tara Matthews, Adam Perer, Inbal Ronen, and Ido Guy. 2012. Diversity among enterprise online communities: collaborating, teaming, and innovating through social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2815–2824.
[42] S. Murugesan. 2007. Understanding Web 2.0. *IT Professional* 9, 4 (July 2007), 34–41. https://doi.org/10.1109/MITP.2007.78

[43] Ted Nelson. 1965. A file structure for the complex, the changing, and the indeterminate. *ACM Annual Conference. Proc. of the 1965 20th National Conference* (1965), 84–100.
[44] Blair Nonnecke and Jenny Preece. 2001. Why lurkers lurk. *AMCIS 2001 Proceedings* (2001), 294. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1733&context=amcis2001
[45] Katherine Panciera, Aaron Halfaker, and Loren Terveen. 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*. ACM, 51–60. http://dl.acm.org/citation.cfm?id=1531682
[46] Patchareeporn Pluempavarn, Niki Panteli, Adam Joinson, Dawn Eubanks, Leon Watts, and James Dove. 2011. Social roles in online communities: Relations and trajectories. In *6th Mediterranean Conference on Information Systems, Nicosia, Cyprus. Retrieved October*, Vol. 4. 2012. http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1042&context=mcis2011
[47] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*. (2011). http://dspace2.flinders.edu.au/xmlui/handle/2328/27165
[48] Jenny Preece. 2000. *Online communities: Designing usability and supporting socialbilty*. John Wiley & Sons, Inc.
[49] Jennifer Preece and Ben Shneiderman. 2009. The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction* 1, 1 (2009), 13–32. http://aisel.aisnet.org/thci/vol1/iss1/5/?utm_source=twitterfeed&utm_medium=twitter
[50] Dana Rotman, Kezia Procita, Derek Hansen, Cynthia Sims Parr, and Jennifer Preece. 2012. Supporting content curation communities: The case of the Encyclopedia of Life. *Journal of the Association for Information Science and Technology* 63, 6 (2012), 1092–1107.
[51] Philipp Singer, Fabian FlÃűck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. Evolution of reddit: from the front page of the internet to a self-referential community?. In *Proceedings of the 23rd international conference on world wide web*. ACM, 517–522.
[52] Yla R. Tausczik and James W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.
[53] Steven JJ Tedjamulia, Douglas L. Dean, David R. Olsen, and Conan C. Albrecht. 2005. Motivating content contributions to online communities: Toward a more comprehensive theory. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*. IEEE, 193b–193b.
[54] Loren Terveen, Will Hill, Brian Amento, David McDonald, and Josh Creter. 1997. PHOAKS: A system for sharing recommendations. *Commun. ACM* 40, 3 (1997), 59–62. http://dl.acm.org/citation.cfm?id=245122
[55] Jason Tsay, Laura Dabbish, and James Herbsleb. 2014. Let's talk about it: evaluating contributions through discussion in GitHub. In *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*. ACM, 144–154.
[56] Christian Wagner. 2004. Wiki: A technology for conversational knowledge management and group collaboration. *Communications of the association for information systems* 13, 1 (2004), 19.
[57] Christian Wagner. 2006. Breaking the knowledge acquisition bottleneck through conversational knowledge management. *Information Resources Management Journal (IRMJ)* 19, 1 (2006), 70–83.
[58] Etienne Wenger, Richard Arnold McDermott, and William Snyder. 2002. *Cultivating communities of practice: A guide to managing knowledge*. Harvard Business Press.
[59] Dennis M. Wilkinson. 2008. Strong regularities in online peer production. In *Proceedings of the 9th ACM conference on Electronic commerce*. ACM, 302–309. http://dl.acm.org/citation.cfm?id=1386837
[60] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 255–264.
[61] Xinchuan Zeng and Tony R. Martinez. 2000. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* 12, 1 (2000), 1–12.
[62] Haiyi Zhu, Robert Kraut, and Aniket Kittur. 2012. Effectiveness of shared leadership in online communities. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 407–416. http://dl.acm.org/citation.cfm?id=2145269

Received April 2018; revised July 2018; accepted September 2018