EVOLUTION OF COMMUNITIES: HOW NEW NETWORKS EMERGE FROM OLD

[1]Umesh

[1]Senior Grade Lecturer
[1]Department of Science,
[1]Government polytechnic, Aurad (B), India

Abstract:

The emergence of new communities is a key topic of study in the social sciences. While much research focuses on how individual social networks contribute to forming a single community, we introduce a fresh, community-centric perspective. Our approach emphasizes that new communities arise within a landscape of pre-existing ones. By analyzing the previous community affiliations of early members, we uncover the process through which communities emerge.

For our analysis, we use Reddit, a platform hosting tens of thousands of user-generated communities. Our dataset spans over a decade, covering user posting activity from Reddit's inception to April 2017. We develop a computational framework to construct genealogy graphs that map relationships between communities. This enables us to conduct the first large-scale analysis of such genealogy structures. Interestingly, fundamental graph properties—such as the number of parent communities and the highest parent connection strength—stabilize quickly, despite the rapid increase in the number of communities over time.

Additionally, we explore the relationship between a community's origins and its future growth. Our findings reveal that strong connections to parent communities correlate with a higher likelihood of future expansion, reinforcing the role of pre-existing structures in the development of new communities. Lastly, at the individual level, we analyze early adopters and their characteristics. We discover that having a diverse engagement history across multiple communities is the strongest predictor of an individual becoming an early member of a newly formed community.

This study has been undertaken to investigate the determinants of stock returns in Karachi Stock Exchange (KSE) using two assets pricing models the classical Capital Asset Pricing Model and Arbitrage Pricing Theory model. To test the CAPM market return is used and macroeconomic variables are used to test the APT. The macroeconomic variables include inflation, oil prices, interest rate and exchange rate. For the very purpose monthly time series data has been arranged from Jan 2010 to Dec 2014. The analytical framework contains**.**

Index Terms - Community Evolution, Social Networks, Genealogy Graphs, Online Communities, Network Analysis

I. INTRODUCTION

The natural inclination of individuals to gather and form groups has led to the continuous emergence of new communities, both in digital and physical spaces. Platforms that enable users to create communities freely—such as Facebook, Reddit, and 4chan—offer valuable insights into this phenomenon. For instance, Figure 1 illustrates the sharp rise in the number of communities since Reddit introduced the ability for users to organize themselves into topic-based groups. This growing trend raises an intriguing question: what are the origins of these newly formed communities?

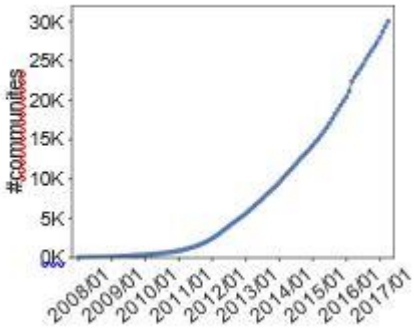

Figure 1: The monthly total of Reddit communities with over 100 members: the number of communities has surged since 2008, when users gained the ability to create their own groups.

In this study, we tackle this question by treating each community as an independent entity, identifying its parent communities, and constructing a genealogy of communities. While numerous studies have explored group formation and community growth within online platforms (Backstrom et al., 2006; Kairam, Wang, & Leskovec, 2012; Kossinets & Watts, 2006; Liben-Nowell & Kleinberg, 2008; Pavlopoulou et al., 2017), most focus on individual communities without considering the broader ecosystem of existing groups. Our approach aligns with research that explains the emergence of organizations by analyzing interactions among a small set of closely related communities (Padgett & Powell, 2012).

For example, Fleming et al. (2007) highlight how academic institutions and industry labs play a role in the rise of high-tech companies in Silicon Valley and Boston. Similarly, our work introduces a computational framework to trace the origins of a community within the landscape of pre-existing ones.

A key insight from our study is that while every new community begins with zero members, it does not emerge in isolation. Instead, its early members bring with them a history of participation in other communities. By examining the past affiliations of these initial members, we can uncover the roots of a new community and map its evolutionary path.

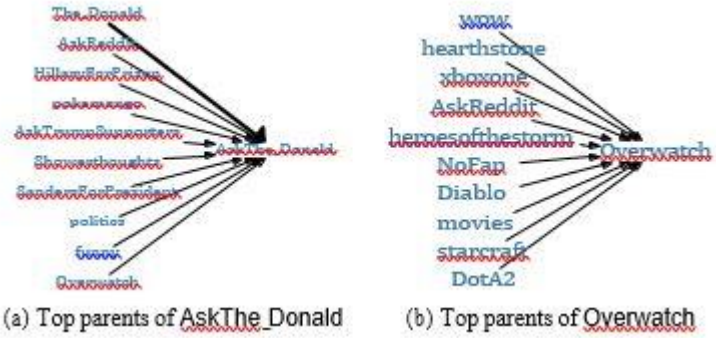

Figure 2: Genealogy graphs illustrating the origins of select Reddit communities based on their first 100 members. A directed edge signifies that some early members of the "child" community were previously part of the "parent" community. The thickness (weight) of an edge reflects the proportion of such members. Node color corresponds to depth within the genealogy graph, with darker shades representing older communities, while node size indicates community membership count.

Figures 2a and 2b highlight the top 10 parent communities of *AskTheDonald* and *Overwatch*, ranked by edge weights. Figure 2c displays a broader genealogy graph tracing connections from two of Reddit's earliest communities (*politics* and *gaming*) to *AskTheDonald* and *Overwatch*. For clarity, only edges with a weight greater than 0.01 (representing at least two members from the parent community) are shown.

Reddit as a Case Study for Community Genealogy

To construct a genealogy of online communities, it is essential to track the activity history of their members. Reddit serves as an ideal platform for this purpose. Figures 2a and 2b illustrate the top 10 parent communities of *AskTheDonald* and *Overwatch*, based on their first 100 members. Unsurprisingly, *AskTheDonald*, a forum where users ask Trump supporters questions, originates from *TheDonald*, a subreddit dedicated to Trump supporters, as well as other politically inclined communities like *HillaryForPrison* and *SandersForPresident*. Similarly, *Overwatch*, a subreddit centered on Blizzard's popular game, stems from gaming communities such as *WoW*, *Hearthstone*, and *Diablo*.

A key difference between these two cases lies in the distribution of influence among parent communities. While *TheDonald* dominates *AskTheDonald's* origins, *Overwatch* lacks a single overwhelmingly dominant parent.

Figure 2c offers a broader perspective, tracing community relationships back to some of Reddit's earliest subreddits, such as *Politics* and *Gaming*. Several key observations emerge:

1. Political and gaming communities form distinct clusters.
2. Political communities exhibit denser and stronger connections, as seen in links between *TheDonald* and *HillaryForPrison*.
3. Some newer subreddits, like *AskReddit* and *Showerthoughts*, have grown to become among the largest despite being created later.

Large-Scale Analysis and Key Findings

This study provides the first large-scale characterization of community genealogy graphs and explores how a community's origins influence its future expansion. Additionally, we examine the traits of early members to better understand the fundamental building blocks of these genealogy networks.

Structure and Key Insights

We begin by reviewing previous research on group formation to provide context. Next, we introduce our dataset, which spans over ten years of Reddit's history.

Our framework for constructing genealogy graphs is based on the first *k* members of a community. By tracking how genealogy graphs evolve as *k* increases from 0 to 100, we observe that:

- The number of parent communities rises ask grows.
- The average influence of individual parent communities decreases, suggesting that the new community becomes less dependent on any single predecessor.

Additionally, despite the rapid growth of new communities on Reddit, the properties of these genealogy graphs stabilize quickly. For example, once a subreddit reaches 100 members, it tends to have approximately 180 parent communities, with the most influential parent contributing around 10% of its early membership.

Predicting Community Growth

We further analyze how a community's genealogy relates to its future expansion. Our findings indicate that genealogy graph data can improve predictions of a community's growth rate, reducing mean squared error by 8.7%. The presence of strong parental ties plays a critical role in determining a community's success, similar to the way political hashtags spread through tightly connected groups.

Characteristics of Early Members

To better understand early adopters at an individual level, we formulate a prediction model. Our analysis reveals that having a diverse membership history across multiple communities is the strongest indicator of becoming an early member of a new subreddit. In contrast, factors such as community feedback and language use are less significant.

This aligns with research on early adopters of new products, which distinguishes between *opinion leaders* and *market mavens*—individuals with broad knowledge about various products and markets. Our study suggests that early members of emerging communities are more akin to market mavens, rather than opinion leaders.

II. RELATED WORKS

Group formation and evolution have long been a central focus in social science research, with scholars such as Lewin (1951) and Coleman (1990) studying the dynamics of how groups emerge and change over time. This paper explores the formation and evolution of online communities, specifically on Reddit, by examining the relationships between new and existing communities rather than focusing on individual user behavior. Traditionally, group formation has been viewed as a diffusion process, where joining a new community is analogous to adopting an innovation. Prior research has demonstrated that the likelihood of an individual joining a community is influenced by the number of their friends already present in that group (Backstrom et al., 2006). Furthermore, the concept of complex contagion, introduced by Centola and Macy (2007), suggests that behaviors such as political engagement require repeated exposure and strong social connections among early adopters. Instead of analyzing group formation at the individual level, this study presents a broader perspective by examining how new communities emerge in relation to existing ones.

Another important aspect of community formation is the distinction between individual identity and community identity. According to social identity theory (Tajfel, 1982), an individual's self-perception is shaped by their group memberships. Previous research has focused on user engagement in multiple online communities and inter-group interactions (Tan & Lee, 2015; Hamilton et al., 2017). However, this study shifts the focus from individual identity to community identity by investigating the origins of communities and their evolution over time. This approach aligns with Astley's (1985) ecological perspective, which differentiates between population ecology and community ecology. Instead of tracing individual user trajectories, the study explores how new communities emerge from existing ones.

To conduct this analysis, the authors utilize a large-scale dataset from Reddit, covering user posts from its inception until April 2017. Reddit initially featured a few default communities, but in 2008, the platform introduced a feature allowing users to create their own subreddits. Each subreddit functions as a distinct community with its own rules and norms. The study focuses on subreddits with at least 100 members to ensure sufficient user activity for meaningful analysis. By constructing genealogy graphs, the researchers trace the emergence of communities over time and examine how they evolve as new users join. Notably, despite the introduction of stricter policies in 2015 that made subreddit creation more challenging, the number of communities continued to grow rapidly, highlighting the resilience of online community formation.

This research contributes to existing literature on online community studies, particularly in areas such as community loyalty and engagement (Hamilton et al., 2017; Kraut et al., 2012), as well as implicit community detection using network structures (Clauset, Newman, & Moore, 2004). By focusing on explicit community formation patterns, the study offers valuable insights into how online communities develop, sustain themselves, and evolve over time. The findings emphasize that community growth persists despite platform restrictions, and the introduction of genealogy graphs provides a novel framework for analyzing the interconnected nature of online communities.

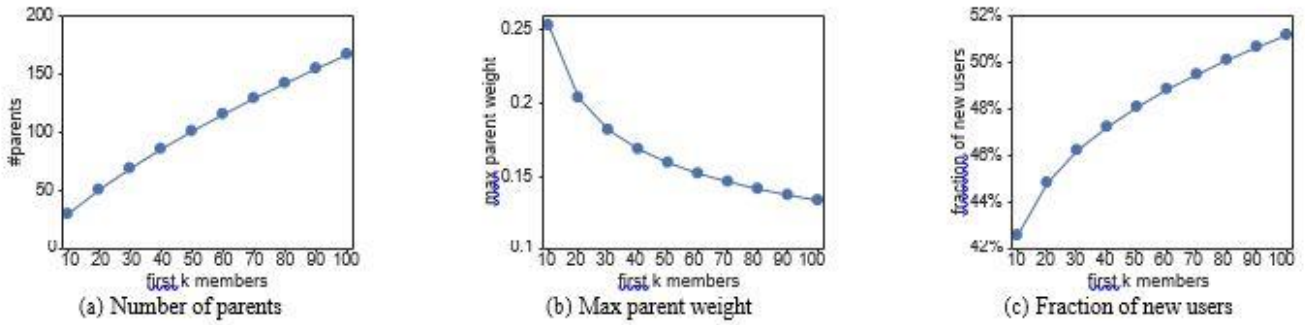

Figure 5: Evolution of genealogy graph properties over time. Each point represents the average corresponding properties of communities created in that month, and error bars represent standard errors. Both the number of parents and max parent weight converge rather quickly despite the rapid increase in the number of communities.

The study examines the evolution of online communities using genealogy graphs, focusing on the emergence of new communities and their relationships with existing ones. Figure 3 illustrates how, as a community grows in membership, the number of parent communities increases while the influence of the dominant parent declines. The fraction of new users, who do not have prior community memberships, also rises over time. These trends persist across different community sizes and creation years, addressing concerns related to Simpson's Paradox (Barbosa et al., 2016). To construct genealogy graphs, the study traces the posting history of early members within a month before they joined a new community. For example, Figure 4 demonstrates the genealogy-building process for "AskTheDonald," where the early members' previous activity determines parent communities and their relative influence. The study defines the weight of parent-child connections as the fraction of early members in the new community who were previously active in a given parent community. As communities evolve from their first 10 to 100 members, additional members contribute to a more diverse genealogy graph, reducing the influence of dominant parent communities. Additionally, new communities attract users without prior participation, including both first-time users and reactivated members who had not posted recently. The findings suggest that while community identities are shaped by early adopters, the emergence of a new community fosters independent growth beyond its initial parent influences. This novel approach to building genealogy graphs provides a foundation for future research in understanding the structural evolution of online communities.

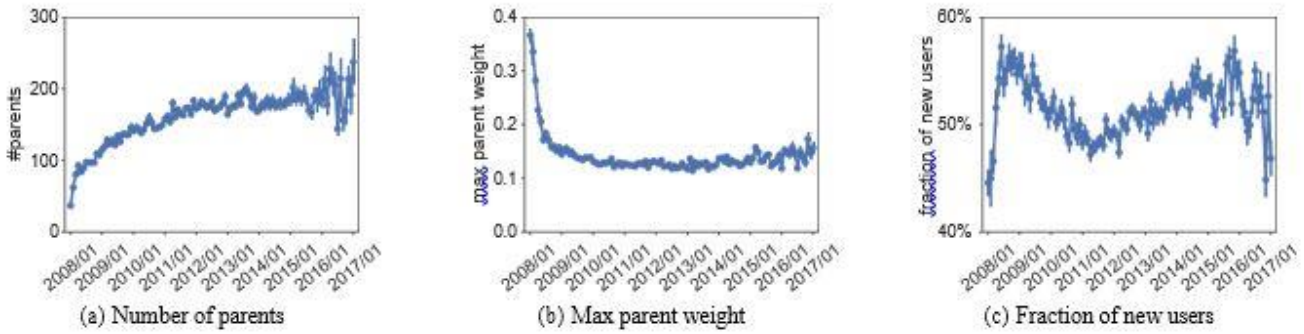

Figure 3 illustrates the relationship between the number of community members (x-axis) and the fundamental properties of genealogy graphs (y-axis). The small error bars represent standard errors. As a community grows in size, the number of parent communities also increases in the genealogy graph, whereas the maximum parent weight gradually decreases. Additionally, the proportion of new users without any recent community affiliations rises. To address potential concerns related to Simpson's Paradox, communities are also grouped based on their year of creation and size. The observed trends remain consistent under all conditions (Barbosa et al., 2016), and this pattern is also reflected in **Figure 5**.

To assess the dominance of the top parent (e.g., *The Donald* for *AskTheDonald* in **Figure 2a**), we introduce the concept of maximum parent weight, defined as the highest value of *wij* for a child community *j*. The number of initial members (*k*) plays a crucial role in this definition. We use small values of *k* to analyze the early formation of a community when its overall size is still relatively small (≤100 members). For accuracy, we use an absolute count of *k* rather than a relative percentage. As *k* increases, a community begins to develop its unique identity. By varying *k* from 10 to 100, we track the evolution of the genealogy graphs.

Figure 4 provides a visual representation of the genealogy-building process for *AskTheDonald*. The timeline highlights the posting sequences of its early members. By analyzing the recent posting activity of *u1* and *u2*, we calculate the weights in the genealogy graph for $k = 10$, assuming that *u3* through *u10* had not posted in the previous month.

Constructing Genealogy Graphs

The fundamental approach to establishing genealogy graphs for communities involves identifying a set of "parent" communities for each newly formed community. We determine these parent communities by analyzing where early members were active just before joining the new community. Essentially, we define a community's origins based on the recent memberships of its early members. Specifically, a new community *j* is assigned parent communities based on the posting history of its first *k* members during the month before they posted in community *j*. **Figure 4** illustrates this for *AskTheDonald* when $k = 10$. The weight of an edge between parent community *i* and child community *j* is determined by the fraction of early members in *j* who were previously active in *i*.

This study represents the first attempt to construct genealogy graphs linking different communities. The genealogical edges are derived from the posting history of early members, meaning that parent-child relationships in these graphs indicate where a new community's members originated. However, it is important to note that these relationships are not necessarily semantic; for instance, in **Figure 2a**, *Overwatch* is considered a "parent" of *AskTheDonald*.

From Zero to One Hundred Members

As the saying goes, "Rome wasn't built in a day," and the same principle applies to online communities—they all start from zero members, regardless of their eventual size. We examine the early development of communities by tracking their genealogy graphs as *k* increases from 10 to 100.

By analyzing how the core properties of genealogy graphs evolve during a community's early growth (see **Figure 3**), we observe that as a community attracts more members, it also accumulates a richer history of past affiliations, leading to an increasing number of parent communities. At the same time, the influence of any single parent community diminishes, as indicated by the declining trend in maximum parent weight. This decline reflects the evolving nature of a new community's identity.

Additionally, a new community serves as an entry point for users who have never been part of any prior community. Over time, we notice an increasing proportion of these entirely new users.

Evolution Over Time

Beyond analyzing the early emergence of individual communities, we also examine how the properties of genealogy graphs change over time as Reddit itself expands. Despite the rapid proliferation of new communities (**Figure 1**), the structural properties of genealogy graphs stabilize relatively quickly. Our discussion primarily focuses on $k = 100$, though similar patterns hold across different values of *k*.

- **Number of Parent Communities:** Over time, the number of parent communities for new communities grows and stabilizes at around 180 (**Figure 5a**). As Reddit expands and users explore more communities (Tan & Lee, 2015), the number of parent communities increases due to members bringing diverse previous affiliations. However, since 2012, this growth has plateaued, suggesting that for communities formed after that point, the first 100 members typically have histories in about 180 existing communities.
- **Max Parent Weight:** This metric, which reflects the dominance of the most influential parent community, declines and stabilizes at 0.1 (**Figure 5b**). With the growing number of communities, new members tend to originate from increasingly diverse backgrounds. Consequently, max parent weight declines from approximately 0.4 to 0.1. Interestingly, this shift happens rapidly—by early 2009, max parent weight had already stabilized at 0.1, indicating that, on average, 10% of a new community's first 100 members share the same parent community. This pattern persists despite the overall number of communities increasing from the hundreds to tens of thousands.
- **Proportion of New Users:** The fraction of brand-new users fluctuates over time (**Figure 5c**). Unlike the other two properties, this metric does not exhibit a clear convergence. Initially, the proportion of new users increased when Reddit first introduced user-created subreddits in early 2008. However, this trend reversed until around 2011, after which the proportion of new users began to rise again. One possible explanation for this increase is the shutdown of the original *reddit.com* main page, which may have prompted some users to explore new communities. However, this alone does not fully account for the continued growth in the fraction of new users between 2011 and 2014.

Predicting Community Growth

Having established the trends in genealogy graphs over time, we now explore how a community's origins relate to its future expansion. We develop a predictive framework to analyze whether a community's early formation influences its long-term growth. Our findings consistently show that strong connections to parent communities correlate with future growth.

Problem Setup

To understand the relationship between a community's origins and its future trajectory, we formulate two prediction tasks:

1. **Growth Classification:** Inspired by previous studies on predicting information cascades (Cheng et al., 2014), we categorize communities based on whether their future size surpasses the median community size (341 members in our dataset). This setup offers two key advantages: (1) It controls for a community's initial size, focusing solely on future growth, and (2) it results in a balanced classification problem, with a baseline accuracy of approximately 50%.
2. **Growth Rate Regression:** For communities that do surpass 341 members, we estimate how long it takes to reach this threshold. Using the formula $log(t341 - t100)$ as the target variable, we frame this as a regression problem, where tk represents the time when the kth member joined. This approach focuses on about half of the communities from the previous classification task.

For both prediction tasks, we assess how much information can be gained by analyzing only the first k members (where k = 10, 20, ..., 90). This allows us to evaluate the impact of additional early members on predictive accuracy.

III. Concluding Discussion

In this study, we examine the formation of communities and provide the first large-scale analysis of genealogy graphs, which trace connections between communities based on the previous affiliations of their early members. Our findings reveal that as a new community takes shape, the number of its predecessor communities rises while the influence of each predecessor diminishes. Interestingly, despite the rapid increase in the number of communities on Reddit, the number of parent communities and their maximum influence tend to stabilize rather quickly. Furthermore, we highlight how a community's origins can effectively predict its future expansion, particularly in terms of growth rate. Our results indicate that strong ties to parent communities are linked to a community's success, supporting the concept of complex contagion, where the spread of behaviors requires dense interconnections among early adopters (Centola and Macy 2007). Additionally, our analysis of early members reveals that individuals with diverse community engagement play a crucial role in a community's growth.

This research represents an initial effort to understand how new communities emerge within the broader context of existing ones. Our observations open several avenues for further investigation. For example, the stabilization of key graph properties over time could be influenced by cognitive constraints on individuals (Dunbar 1992) or by specific structural requirements for new communities to take shape—such as the finding that approximately 10% of early members share prior experience in an existing community. Understanding the factors driving this stabilization remains an open question, as does determining whether similar patterns occur on other platforms or under different definitions of membership.

Moreover, the concept of genealogy graphs offers valuable insights into community development and warrants further exploration. While we focus on a fixed number of early members to study the emergence process, community formation is inherently dynamic, varying across user bases and timeframes. Future research could refine the definition of early members to accommodate different types of communities and even develop a universal framework for identifying various stages of community development. Another compelling avenue of inquiry is the influence of community origins and early member characteristics across diverse community types, such as political versus gaming groups or common-bond versus common-identity groups (Prentice, Miller, and Lightdale 1994; Ren, Kraut, and Kiesler 2007). Lastly, the contributions of individual members in a genealogy graph are not necessarily equal. The nature of the relationship between parent and child communities may depend on the motivations behind members joining the new community. Investigating these connections in greater detail could lead to the development of genealogy graphs enriched with more nuanced relational data.

References

[1] Abratt, R.; Nel, D.; and Nezer, C. 1995. Role of the Market Maven in Retailing: A General Marketplace Influencer. *Journal of Busi- ness and Psychology* 10(1):31–55.

[2] Aral, S.; Muchnik, L.; and Sundararajan, A. 2009. Distinguish- ing Influence-based Contagion from Homophily-driven Diffusion in Dynamic Networks. *Proceedings of the National Academy of Sciences* 106(51):21544–21549.

[3] Astley, W. G. 1985. The Two Ecologies: Population and Com- munity Perspectives on Organizational Evolution. *Administrative Science Quaterly* 30(2):224–241.

[4] Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *KDD*.

[5] Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an Influencer: Quantifying Influence on Twitter. In *WSDM*.

[6] Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The Role of Social Networks in Information Diffusion. In *WWW*.

[7] Barbosa, S.; Cosley, D.; Sharma, A.; and Cesar Jr, R. M. 2016. Averaging gone wrong: Using time-aware analyses to better under- stand behavior. In *WWW*.

[8] Baumgarten, S. A. 1975. The Innovative Communicator in the Diffusion Process. *Journal of Marketing Research* 12(1):12–16.

[9] Centola, D., and Macy, M. 2007. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology* 113(3):702–734.

[10] Centola, D. 2010. The Spread of Behavior in an Online Social Network Experiment. *Science* 329(5996):1194–1197.

[11] Cheng, J.; Adamic, L. A.; Dow, P. A.; Kleinberg, J.; and Leskovec,

J. 2014. Can Cascades be Predicted? In *WWW*.

[12] Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding Community Structure in Very Large Networks. *Physical Review E S* 70(066111).

[13] Clauset, A.; Shalizi, C. R.; and Newman, M. E. 2009. Power-law distributions in empirical data. *SIAM review* 51(4):661–703.

[14] Coleman, J. S. 1990. Foundations of Social Theory.

[15] Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *WWW*.

[16] Ducheneaut, N.; Yee, N.; Nickell, E.; and Moore, R. J. 2007. The Life and Death of Online Gaming Communities. In *CHI*.