Science, AskScience, and BadScience:
On the Coexistence of Highly Related Communities

Jack Hessel and **Chenhao Tan** and **Lillian Lee**

Department of Computer Science
Cornell University
{jhessel, chenhao, llee}@cs.cornell.edu

Abstract

When large social-media platforms allow users to easily form and self-organize into interest groups, highly related communities can arise. For example, the Reddit site hosts not just a group called food, but also HealthyFood, foodhacks, foodporn, and cooking, among others. Are these highly related communities created for similar classes of reasons (e.g., *true* to distinguish one as a better community and *advice* to focus on helping fellow members)? How do users allocate attention between such close alternatives when they are available or emerge over time? Are there different types of relations between close alternatives such as sharing many users vs. a new community drawing away members of an older one vs. a splinter group failing to cohere into a viable separate community? We investigate the interactions between highly related communities using data from reddit.com consisting of 975M posts and comments spanning an 8-year period. We identify a set of typical affixes that users adopt to create highly related communities and build a taxonomy of affixes. One interesting finding regarding users' behavior is: after a newer community is created, for several types of highly-related community pairs, users that engage in a newer community tend to be *more active* in their original community than users that do not explore, even when controlling for previous level of engagement.

Introduction

Social networks are in constant flux, with new communities forming and old communities dying over time. On websites such as Facebook and Reddit, users have complete freedom to create communities at their own discretion. This has led to a very large number of communities arising organically from user initiative. The motivations for creating new groups vary. One reason is to create divisions that satisfy the need to better organize discussions; in fact, community design theory argues that "a growing Web community needs subdivisions which might be represented as towns, neighborhoods, topics, categories, conferences, or channels, depending on your metaphor" (Kim 2000; Jones and Rafaeli 2010). Or, new groups can develop because of religious, political, or other schisms; online examples include groups whose very names attempt to connote superiority to others, e.g., the subreddits trueatheism vs.

Table 1: The 10 most common Reddit group-name affixes.

Affix	Example	# Pairs
s	auto, autos	63
porn	space, spaceporn	26
circlejerk	hiphop, hiphopcirclejerk	23
ask	science, askscience	21
shitty	ideas, shittyideas	17
music	running, runningmusic	17
help	tech, techhelp	11
2	dota, dota2	9
true	atheism, trueatheism	9
learn	math, learnmath	9

atheism. Other reasons surely exist. The tremendous reach of modern social media provides researchers much greater data to examine these social processes at scale.

An interesting and frequently occurring version of the group creation process is that a new concept or culture may gain in popularity and, in a meme-like fashion, draw users to create a new community by using that concept as *an affix*[1] of their community name. For example, on Facebook, after the creation of the OMG Confessions group, anonymous confession pages with names combining a college with the word *confession* or *confessional* proliferated to the degree that one can now find a confession page for almost every university campus. (Birnholtz, Merola, and Paul (2015) examined what kind of questions people ask on such pages.) Table 1 shows some examples from Reddit: the second column shows pairs of subcommunities where the name of one is a modified form of the other (ignore the third column for now).

A whimsical Reddit example is that the affix *random_acts_of_*, as a general name that means people sending free things altruistically to others that are in need, has attracted users to create random_acts_of_pizza, random_acts_of_amazon, and random_acts_of_books, etc. (Althoff, Danescu-Niculescu-Mizil, and Jurafsky (2014) used random_acts_of_pizza to study what are the most effective ways to ask for a favor).

In this work, we investigate highly related communities

---

[1] An affix is either a prefix or a suffix.

that are based on affixes. An understanding of these highly related communities may help community organizers identify subtopics in a community and create an appropriate subdivision to cultivate focused discussions, or monitor subgroups that potentially feel marginalized or underserved, and decide whether to change community norms or create a dedicated community for that subgroup.

Despite the ubiquity and importance of such affixes, little is known about canonical affixes and the activity in the resultant highly related communities. For instance, are neighborhoods, topics, and channels enough to capture all possible affixes? Are there classes of affixes that are generally applicable?

Once a highly related communities is created, how does it interact with the existing community? Will it overtake it? Will the two share the same userbase? Perhaps different affixes behave in different ways. For instance, communities that focus on a specific geographic location (e.g., those that are created from the *canada* affix) may not necessarily attract users in the existing community, while affixes such as *true* may end up drawing many away. One of our goals is to analyze user behavior in the existing community behave *after they participate in the new community*.

Organization and contributions. In this paper, we construct a dataset from Reddit and present the first large-scale study on the coexistence of highly related communities. Details about the dataset are introduced in "Dataset Description".

Our first contribution is to characterize the space of *affixes*. We build a taxonomy of common affixes that users adopt to create highly related communities. For instance, we identify a category of "parody" affixes (*circlejerk*, *shitty*, *funny*, *lol*, *bad*). This category generally shares the same user base with its corresponding unaffixed community. On the other hand, we identify a category of "derivative" affixes (*meta*, *anti*, *srs*, *post*, *ex*) that likely attract different user bases. Surprisingly, a non-trivial fraction of affixed communities exist before the unaffixed ones. Also, an interesting class of *spinoff* communities arises where early participants in the new community are from the existing community.

Our second contribution is to introduce a framework for analyzing users who try out spinoff communities (dubbed "explorers") and comparing them to "nonexplorers" who never leave the original subreddit. We make the surprising observation that in multiple classes of affixes, users who explore spinoff communities are *more active in the original communities after exploring* when compared to similarly active users who never tried the alternative. This resonates with the findings in Tan and Lee (2015) that users who "wander" to different (potentially completely unrelated) groups tend to stay active longer on the site as a whole. Our observations suggest that spinoff communities generally serve a complementary rather than competitive role in multicommunity settings.

In the end, we summarize related work and offer some concluding thoughts.

Table 2: Summary statistics for our Reddit corpus. Posts are from Tan and Lee (2015) and include all posts on Reddit from its inception in 2006 to February, 2014. Associated comments of these posts are added from Jason Baumgartner's comment dataset until Nov 2014.

Data type	count
Subreddits	5,692
Posts	88M
Comments	887.5M

Dataset Description

Our starting point for understanding highly related communities, affixes, spinoffs, nonexplorers, and explorers is an examination of *topically related communities*. As such, we compile a dataset from `reddit.com`, a site where users are allowed to create communities called subreddits at their discretion. Users can name the subreddits that they create so that like-minded people can identify them effectively. As a result of unmoderated creation and limitless naming possibilities, there are a wide variety of subreddits on Reddit, e.g., funny, worldnews, politics, IAmA, todayilearned, etc. On these subreddits, users submit link-based posts or text-based posts, comment on others' posts, and up/down vote posts and comments. We construct a dataset that includes all activities on Reddit from its inception until 2014, an 8-year period, by combining two data sources: a post dataset that was organized in Tan and Lee (2015), and all comments data extracted by Jason Baumgartner.[2] We focus on communities that are active with a reasonable number of users. Specifically, we require all communities to include at least 300 unique users that made posts. This left us with just under 5.7K communities. Table 2 presents basic statistics of this dataset.[3] The metadata for Reddit conversation trees that we used here is available for download.[4]

As discussed in the introduction, user-defined subreddit names are an important indicator of relationships between highly related communities (e.g., food vs. HealthyFood). We first retrieve all possible pairs of communities where one community name is the other's suffix or prefix, ignoring case (food is the suffix of HealthyFood, ignoring case). We refer to the difference between the names in a pair as the *affix*. For instance, *healthy* is the affix in the pair food vs. HealthyFood. There are around 4K such pairs over our dataset.

Using common affixes as a starting point allows us to discuss the the space of possible highly related communities. For example, this framing allows us to make statistical observations about all pairs with *healthy* or *true* as affixes. Note that we omit some interesting highly related communities pairs by focusing on affixed pairs. One example is TwoX-

---

Chromosomes, a very popular "subreddit ... intended for women's perspectives," and TrollXChromosomes, its satirical counterpart.

Identifying topically related communities. Unsurprisingly, not all pairs of communities identified through affixes are actually highly related communities. An example is "ru" and "rum;" the first one is a Russian community while the second one is about the liquor. In order to quantify subreddit similarity, we compute the content similarity between pairs of communities. As suggested in Singer et al. (2014), subreddits can focus on text posts in addition to link-based posts. Therefore, we employ a method that can account for either link-dominant or text-dominant subreddits. Specifically, we use Jaccard similarity between the set of links to capture similarity based on links,[5] and use Jensen-Shannon divergence between topic distributions (derived from a topic model trained on 6.6M text posts) to capture similarity based on text following Hessel et al. (2015). Since these two metrics are not comparable by raw value, we compute the full background distribution based on all 1.62M possible pairs of the 5.7K communities in our dataset and compute the percentile of each affix pair in each distribution.

We consider a pair of communities to be topically related if either link similarity is above the 90th percentile *or* topical similarity based on text is above the 90th percentile. Accounting for our definition of topical similarity yields just over 1.7K pairs from our original set of 4K.

The last step of our preprocessing is to identify generalizable affixes that are commonly used in these highly related communities. We count the frequency of affixes and keep affixes that occur at least three times, so that all affixes in the final dataset carry a general meaning (it is not possible to make general statements about affixes that only occur once). This step brings us to 99 affixes and 572 pairs of highly related communities distributed between them.

Characterizing affixes

The goal of this section is to explore the types of canonical affixes users on Reddit utilize. To accomplish this exploration, we first build a taxonomy of common affixes to better understand their basic properties and relationships. Next, we explore the temporal characteristics of the pairs. In general, we observe an accelerating culture of creating highly related communities, meaning that highly related communities are being created at increasing rates. We also observe that, in most cases, the affixed community in a pair was created after the unaffixed one, even though there is a non-trivial fraction that went the other way, e.g., ukpolitics and uspolitics both existed before politics. We further explore whether the newer community "overtakes" the older one in popularity. We then offer rationale to explain the surprising finding that a *quarter of the newer communities are more active.* The final characteristic that we examine is whether the newer community actually shares a user base with the older one, at least when the new one is forming. Despite the high similarity both in community name and in content, almost half

---

[5]Jaccard similarity is defined as $\frac{A \cap B}{A \cup B}$, where A and B are the set of links from two subreddits respectively.

Table 3: A taxonomy of affixes.

| Adjective-like	
"better"	true, plus
"parody"	circlejerk, shitty, funny, lol, bad
"derivative"	post, ex, meta, anti, srs
"genre"	classic, fantasy, indie, folk, casual, dirty, classic, metal, academic, 90s, free, social
"nsfw"	nsfw_, nsfw, asian, trees, gonewild, gw, r4r, tree
Verb-like	
"learning, improvement"	ask, help, learn, advice, hacks, stop
"action"	exchange, randomactsof, trade, trades, classifieds, market, swap, random_acts_of_, requests, invites, builds, making, mining, craft
Noun-like	
"place"	uk, reddit, chicago, us, dc, steam, canada, american, boston, android, online, web
"medium"	porn, pics, music, memes, videos, vids, comics, apps, games, gaming, game
"subject"	science, news, dev, servers, tech, tv, guns, recipes, city, u, college, man, girls
Minor	
"equivalent, competition"	s, al, ing, the, alternative
"generation"	2, 3, 4, 5
"modifier"	ism, n, an

of newer subreddits in pairs are not, in fact, born out of their older partners.

The space of affixes

In order to achieve a basic understanding of what canonical affixes that users adopt to create new communities, we first build a taxonomy of the 99 affixes from the dataset section in Table 3.

We start with a coarse structure based on part-of-speech. Among the adjective-like, the largest category is based on "genre", e.g., rock vs. classicrock. Some other very interesting classes also arise: "better" that indicate a certain level of superiority (e.g., atheism vs. trueatheism), communities dedicated to "parody" where users are likely aware of the culture in the unaffixed one (e.g., history vs. badhistory), and "derivative" that probably attracts very different audience (e.g., war vs. antiwar). In fact, *anti* and *meta* can be recursive, e.g., jokes, antijokes and antiantijokes.

Among the verb-like affixes, a class of self-improvement or learning community exists, e.g., programming vs. learnprogramming. In "actions", there are many exchange related affixes, including *trades* (e.g., pokemon vs. pokemon-
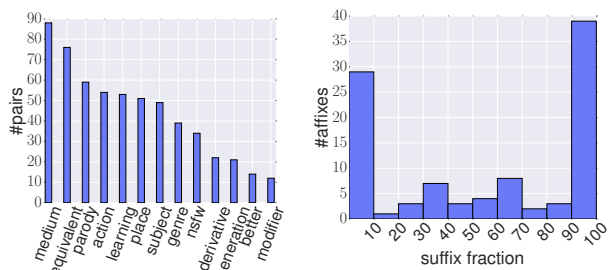
(a) #Pairs by category.

(b) Histogram of fraction as suffixes for all affixes.

Figure 1: (a) Medium is the most frequent affix, while modifier is the least. (b) Two distinct types of affixes exist: suffix-dominant and prefix-dominant.

(a) Average gap for pairs grouped by the creation year of the older community.

(b) Histogram of fractions that the affixed community was created later.

Figure 2: (a) The newer related community is created with shorter and shorter gap. (b) For most affixes, the affixed community was created later, though there are many counterexamples.

trades) and *swap* (e.g., scotch vs. scotchswap). Altruistic behavior signified by *random_acts_of_* (e.g., pizza vs. random_acts_of_pizza) has been studied specifically in (Althoff, Danescu-Niculescu-Mizil, and Jurafsky 2014).

The noun-like affixes closely match the conceived metaphor of splitting space in community design theory (Kim 2000). Indeed, we see a group of affixes based on "place", such as *uk* (e.g., politics vs. ukpolitics). "Medium" is another common category including *videos* (e.g., cat vs. catvideos). The last one is based on "subject", such as *recipes* (e.g., vegan vs. veganrecipes). Noun-like affixes are probably used to encourage better discussions. These communities do not necessarily share similar users, i.e., people who are interested in veganrecipes may not be vegans; people who are invested in ukpolitics may not care about politics in general.

Surprisingly, there is a class of relatively minor changes that can cause community pairs to differ significantly. An example of *ism* is vegetarian vs. vegetarianism, which align in topic but likely attract very different group of people Another interesting class is "equivalent"; this class of communities can even be in competition because they are about exactly the same thing. One example is wallpaper vs. wallpapers: these two subreddits have indistinguishable content and thousands of members each, yet the moderator sets are disjoint, and each hesitates to mention each other despite each having extensive lists of related subreddits. In cases like this, the newer community may be created without knowing about the older one. Note that it is not always the case: Politic was created because some users do not like the rules in politics.

Although some decisions in this taxonomy are arbitrary, we consider it useful and meaningful to get an overall sense of possible affixes. All affixes we consider seem to be generalizable changes that one can make with some community name to obtain another community name.

Frequency of affixes. Next, we examine the frequency of affixes. Table 1 presents the 10 most common affixes and Figure 1a shows the frequency by category. The most common affix is simply the character *s*, which suggests that it is perhaps common for "redundant" communities to be created
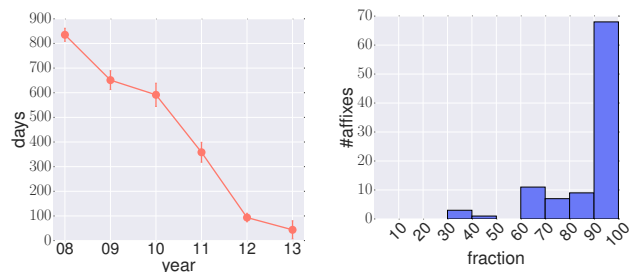
without knowledge of the other. The most common category is "medium" (named for the *medium* of the content, e.g., *pics*, *vids*, etc.) with 88 pairs, while the least one is "modifier" with 12 pairs. There is also quite some diversity within each category. For instance, one interesting observation is that although *porn* and *pics* both fall in "medium" and indicate a related picture driven community, *porn* has more than 4 times more highly related communities (33 vs 7).

Position of affixes. It remains a question how these affixes are used in founding highly related communities. In particular, an affix can be either a suffix or a prefix. If we compute the fraction of being used as suffixes for each affix, as shown in Figure 1b, two classes of affixes emerge: suffix-dominant and prefix-dominant. Overall, "generation", "medium", and "modifier" tend to be used as suffixes, while "genre", "derivative", and "place" are usually used as prefixes. "parody" and "nsfw" can be used either way, for example, *funny* in videos vs. funnyvideos and Guildwars2 vs. guildwars2funny.

Temporal Relationships within Pairs

It is always possible to determine which community in a pair was created earlier. The first characteristic that we examine is the gap between the creation time of two communities in a pair. The overall average gap is 749 days since 2008, when users on Reddit were first allowed to create their own communities. If we compute the average gap grouped by the creation year of the older community, in Figure 2a we see a consistent trend that the newer community is created with a shorter and shorter gap over time. This suggests that there may be an accelerating culture of creating highly related communities over time, or that as there are more users in on Reddit, affixed communities arise more quickly.

For most affixes, the community with the affix was newer. We further examine whether the newer community within a pair is the affixed one. This is indeed the case in 86% of our pairs. However, if we change our focus from pairs to affixes, we find that for 33% of the affixes, there was at least an instance where the affixed version was actually *created*

before the original (see Figure 2b).

We take a closer look at these pairs in which the affixed member exists before the unaffixed one. The four affixes for which the affixed version of the community more often exists first are *ing*, *al*, *ism* and *s*; these are mostly in the "equivalent/competition" class in Table 3. The nature of these affixes again suggests that communities are often created without knowing the existence of the older one. As a result, we observe phenomena like different communities focusing on exactly the same thing (e.g., wallpaper vs. wallpapers) or two communities eventually deciding to explicitly merge into one (e.g., wedding vs. weddings). Communities with different foci but similar names might also fall into this category. For example, vegetarianism and vegetarian may attract quite different groups of people. Also, these four affixes do not cover all possible cases where the affixed was created earlier. For instance, twincitiessocial was created before twincities.

## Does the New Overtake the Old?

Another important characteristic is how active the newer community is compared to the older one after its inception. **Newer communities tend to be less active, but in a quarter of pairs, the newer one is more active.** We compute the log ratio in activity level (the total number of comments plus the total number of posts) between the newer community and the older community with add-one smoothing, only considering actions *after the newer community was created* so that we compare pairs during the same time period. According to this metric, a positive value means more activity in the newer community and a negative value means less activity in the newer community. Figure 3a demonstrates there is a trend that affixed versions of communities tend to be less popular. The mean log ratio is -2.0, which suggests that new community is usually 13.5% as active as the older one. However, a nontrivial fraction of newer communities (25.7%) are more active.

A closer look at the more active newer communities. It's somewhat surprising that 25.7% of newer communities overtake their established counterparts. Why does this occur? Figure 3 presents examples of possible reasons that the younger community might surpass its older counterpart.

The first reason is that the affix is a "magic" combination and the new community naturally appeals to more people. One example is writers vs. fantasywriters. As soon as fantasywriters was created, its activity level was more than 7 times as much as that in writers. Here are top 3 affixes that consistently lead to more activity: the (e.g., stopgirl vs. thestopgirl), ex (e.g., mormon vs. exmormon), steam (e.g., deals vs. steamdeals).

Second, the newer community may be "equivalent" to the older one and the newer one may win the competition. For example, in the case of auto vs. Autos, it took Autos a while to exceed the activity level in auto, but Autos is now much more popular (see Figure 3b).

Third, the newer community may actually be the non-modified (14% of pairs have this property, as we earlier observed) and the newer one might achieve popularity because it is more general. For instance, politics is more popular than
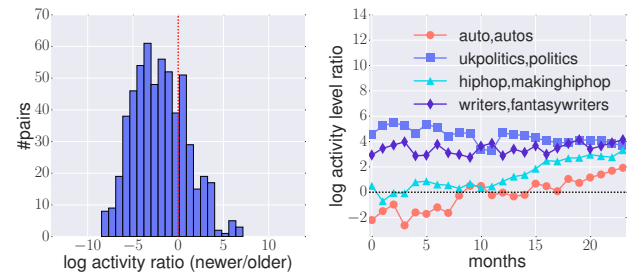
(a) Histogram of log activity level ratio between the newer community and the old one.

(b) Case study on pairs in which the newer one has more activity, where activity is binned on a month-to-month basis.

Figure 3: (a) The older community tend to have a higher level of activity. (b) Examples of different reasons that the newer one can have more activity. It shows how the log activity level ratio changes over time since the newer one was created in the first two years.
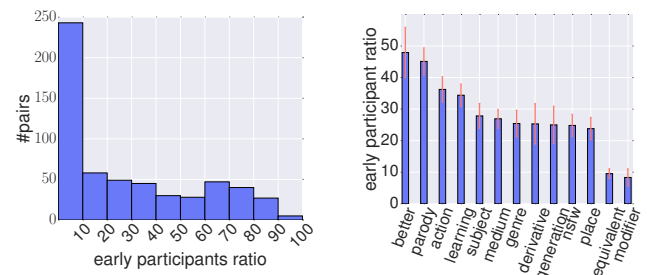
(a) Histogram of the fraction of early participants in the new community that were from the old community over all pairs.

(b) Average fraction of early participants in the new community that were from the old community sorted by categories.

Figure 4: (a) Surprisingly, the majority of highly related communities do not share more than 10% of early participants. (b) "Better" has the highest average early participant ratio, while "modifier" has the lowest.

ukpolitics despite the later's earlier founding. In this case, as soon as politics was created, its activity level exceeded ukpolitics.

The fourth and relatively rare reason is that the older one may have a large competitor, in other words, the newer one may originate from an even bigger community than the older one. An example is hiphop vs. makinghiphop. makinghiphop started at a similar size as hiphop but exceeded hiphop significantly later. Although hiphop and makinghiphop are both active, there is a much larger hiphop-related community on Reddit, hiphophead. makinghiphop might actually originate from hiphophead instead of hiphop.

Where are early participants in the new communities from?

The last reason in the above discussion leads to a natural question: where are the participants in the newer community from? Are they from the older one in a pair? This

question is difficult to answer, as a subreddit may establish its own identity and unique audience over time, even if it was born out of an existing community. If we simply look at the overlap between two communities over all users, we may mistakenly believe that they have never shared the user base as a result of a large number of later users. We thus focus on the first n participants in the newer community (the *early participants*) and compute the fraction of them that were also members of the older community. A user is considered a member of the old community if they made any action in the old community within the last 30 days prior to interacting with the new community. We refer to this metric as "early participant fraction". While we present results for $n = 100$, similar results hold for different n.

Almost half of highly related communities do not really share early participants. As shown in Figure 4a, surprisingly, the majority of newer subreddits in highly related communities pairs are not founded by members of the older community. Although a pair of communities might have similar names, it's common that they are founded by different users. In fact, there are only 7 of the first 100 participants in makinghiphop that were members of hiphop.

Figure 4b presents the average early participant ratio of all categories in Table 3. It shows that "better", "parody", "action" and "learning" usually attract members from the older community. It also partly demonstrates why we obtain such a low average early participant ration. "equivalent" and "modifier" are more than likely to attract completely different participants, e.g., vegetarian vs. vegetarianism. We also notice significant differences even within one category. One notable example is *meta* (65.8% from the original community) vs. *ex* (1% from the original community).

From Highly related communities to Spinoffs

Thusfar, we have explored the complex space of possible affixes, and the highly related communities that are created through them. We find that a non-trivial fraction of the new communities were not the affixed ones, or did not share the same user base of the new one. For these pairs, it is unclear that the new community is a subdivision of the old one, or that users in the existing community affected by the new one's presence. In order to better understand how users in the existing community may behave *after exploring the new community,* we will focus on a subset of highly related communities called *spinoffs* in the remainder of this work.

Spinoffs: Substitutions or Complements?

We now formally define *spinoff* communities. First: the newer of the two pairs in a highly related community is a *spinoff* if it satisfies the following properties: 1) more than 10% of the first 100 early participants in the newer community are members of the older community; and 2) the newer community is the affixed one, so that it is likely to represent a specialization or some other topic of interest. We will sometimes refer to a pair of highly related communities that contain a spinoff as a *spinoff pair.*

In this section, we investigate how a user's behavior within the older subreddit is affected once they try out the
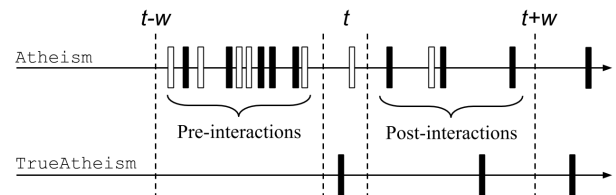

Figure 5: Schematic of the exploration experiment setup. TrueAtheism is a spinoff of Atheism, and the activity of two users is shown over time. Each box represents an interaction. With respect to the two subreddits shown, the dark user is an explorer, and the light user is a nonexplorer. Time t is the time of the black user's first interaction with the spinoff subreddit. Here, the number of pre-interactions for both the dark and light users is 5. The dark user has 3 post-interactions, whereas the light user only has one.

newer spinoff: do such users get "distracted" by the new one, or does the new subcommunity complement the old one? Phrased differently, do users tend to decrease, increase, or not change their activity levels in the original community after trying the spinoff?

Surprisingly, we find that users who explore the spinoff generally become *more* active in the *original* community. Furthermore, with respect to the taxonomy we developed in Table 3, the magnitude of this trend *depends on the type of affix*: larger in "action", "better", and "parody", smaller in "medium", and *negative* in "nsfw". Finally, it seems that this complementary effect is more prominent for users with lower activity level, although there is less data to compare users with different activity level, and results may vary depending on specific pairs.

Disclaimer: *we do not make any claims of causality given the observational nature of our dataset.*

Experiment setup

To understand users' behavior in the original community *after* participating in the spinoff community, we propose an experiment framework by first pairing an explorer and a "similar" nonexplorer in the original community. After identifying this pair of users, we compare their behavior pattern after the explorer first participated in the spinoff community, as illustrated in Figure 5.

Specifically, for each spinoff pair (e.g., Atheism vs. TrueAtheism in Figure 5), we define *explorers* as users who were active in the original community in a window before their first participation in the spinoff community.[6] The darker user in Figure 5 is an example. We denote the time of her first interactions in the spinoff community as t, and refer to her interactions in the original community from $t - w$ to t as *pre-interactions* and her interaction in the original community from t to $t + w$ as *post-interactions*. We consider users with at least 5 pre-interactions to ensure that they were indeed active in the original community.

---

[6]Participation and being active are both defined as either posting or commenting.
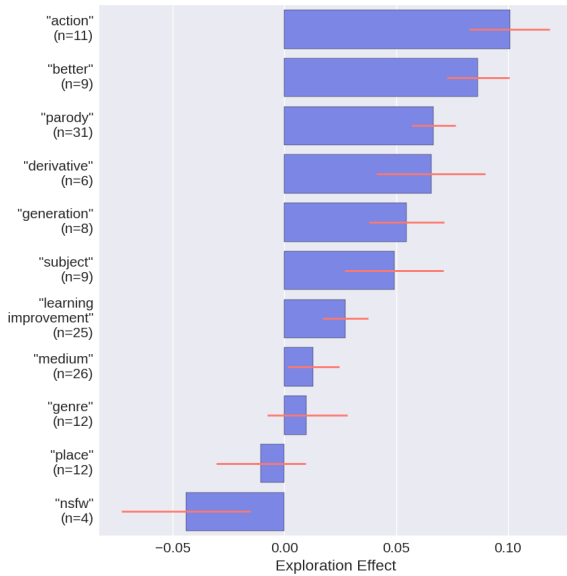
(a) Exploration effect by category (n is the number of sampled pairs, only $n \geq 4$ is shown)
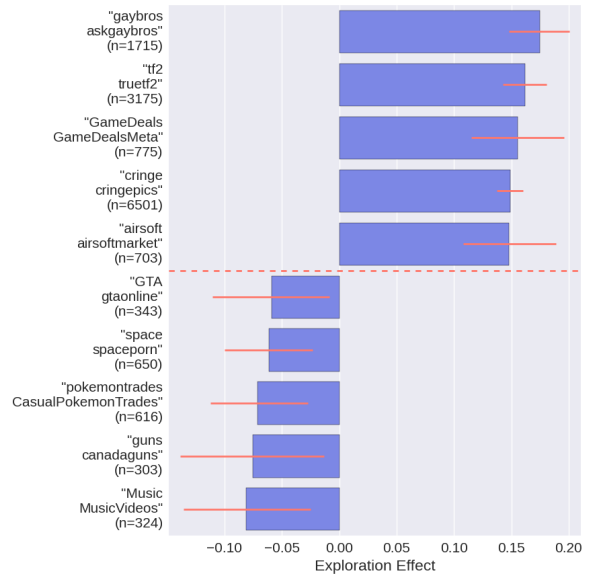
(b) Top/bottom 5 exploration effect by pair (n is the number of sampled users)

Figure 6: Difference between explorers and nonexplorers in the fraction of users that become more active in post-interactions compared to pre-interactions. Larger values indicate more activity from explorers. (a) categories from our taxonomy and (b) specific pairs. Error bars represent 95% CIs.

A straightforward metric to compute is simply the ratio between the number of post-interactions and the number of pre-interactions for each user. However, this is problematic because we require users to have at least 5 pre-interactions but have no constraints on post-interactions. This causes our sample to be biased towards users with more pre-interactions than post-interactions.

To address this concern, for each exploring user u_e, we sample a *similarly active* user u_{ne} in the original community who *never* interacts with the spinoff community. We call this user a "*nonexplorer*". The rough idea is demonstrated by the light user in Figure 5, who had a similar number of pre-interactions and made a post in the original community around t so that we know she was still active. The details of this sampling process are given in the appendix.

Metric: exploration effect. After we identify explorers and matching nonexplorers, we compute the fraction of explorers who have more post-interactions than pre-interactions and that of nonexplorers. We take the difference between these two fractions and call it the "exploration effect" (see Equation 2, in the appendix). Higher values of this quantity indicate that u_e was more active in the original community than the nonexplorer user u_{ne}. We use the macro average to aggregate results from different spinoff pairs as the number of explorers vary between pairs.

The only parameter in our framework is w. Since our primary objective of interest in this work is the effect of the interaction with the spinoff community, we choose a relatively small window (30 days) to mitigate confounding factors that may affect user behavior over time and the dynamic

nature of online communities (Danescu-Niculescu-Mizil et al. 2013; Backstrom et al. 2006; Ducheneaut et al. 2007; Kairam, Wang, and Leskovec 2012; Kumar, Novak, and Tomkins 2010). Our results are robust to reasonable changes in w (e.g., $w = 20$ days produces very similar results).

### More active after exploring the spinoff community

We now apply this framework and examine how explorers behave in general. Surprisingly, we find that explorers are relatively more active compared to nonexplorers, i.e., the exploration effect is generally positive. We then further split explorers based on their activity level and study how our observation differ depending on activity level.

Comparisons across categories. Figure 6a presents exploration effect results for categories in our taxonomy in Table 3.[7] Somewhat counterintuitively, we find that for most spinoffs, users who explore become *more* active in the original subreddit after exploring, compared to similarly active users who never interacted in the new community (see Figure 6a).

Interestingly, the magnitude of this result varies based on the spinoff pair considered. We observe that "action" explorers are around 10% more likely to increase their activity after exploring, for example. "place" explorers, on the other hand, are roughly 2% less likely to increase their activity.

Our possible explanation for this observation is that users who explore to "action" communities are often seeking to

---

[7]Results are reported for categories with more than 4 spinoff pairs.
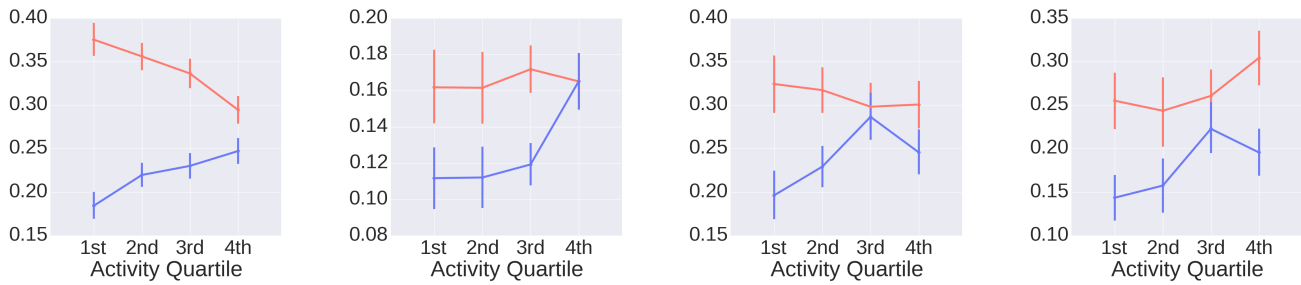
(a) AskReddit vs TrueAskReddit; (n=8816 user pairs) (b) Science vs AskScience; (n=5516 user pairs) (c) Android vs androidapps; (n=2951 user pairs) (d) apple vs applehelp; (n=2221 user pairs)

Figure 7: Several examples of **explorer** and **nonexplorer** activity levels (with 95% CIs) split into quartiles by pre-activity. The x-axis is pre-interaction quartile, and the y-axis is the proportion of users whose post-interactions exceeded their pre-interactions. In all cases, explorers tend to have greater post-interaction levels than nonexplorers, reflective of the results from the previous section. These plots are meant to highlight the complex relationships between activity level and activity rates. We observe many statistically significant differences, but note that each spinoff community pair's behavior in this regard appears to be unique. In the first three pairs, we do see that explorers with the highest pre-activity level present a smaller difference from nonexplorers.

actively engage with a topic in a fashion above and beyond simple discussion. For example, the subreddit Bitcoin (which focuses on high-level discussions of the cryptocurrency) and its spinoff pair BitcoinMining (which focuses on lower-level issues, e.g., hardware useful for mining Bitcoins) exhibits a difference in interaction ratio of roughly 10%. If a user explores in BitcoinMining from Bitcoin, this is likely a strong indication of their interest in digging deeper into the topic itself. It's possible that viewing Bitcoin through the perspective of BitcoinMining increases overall engagement with the topic, at least in the short term.

In contrast, exploring in "place" subreddits does not result in increased home activity nearly as often. For example, Bitcoin has another spinoff pair, BitcoinUK, that has an exploration effect of roughly zero. We have previously seen in Figure 4b that "place" spinoffs share relatively few early participants with their parent communities. Taken together, these observations suggest that that users seeking place-specific communities are not necessarily interested in engaging more deeply with the topic, so much as *who* they discuss the topic with or *how* the topic affects them.

A closer look at the pairs. Figure 6b presents the top and bottom 5 pairs in terms of exploration effect. It further demonstrates how our results may vary across different pairs. All these 5 bottom pairs present significantly negative exploration effect, which shows that it is not always the case that explorers are more active.

Looking at the bottom 5, it partly supported our above discussion regarding places. Indeed, in "place" related pairs, gtaonline distracted people from gta and so did canadaguns for guns. Among the top 5, there is an even spread among several categories including "learning" (gaybros vs. askgaybros), "action" (airsoft vs. airsoftmarket), and "medium" (cringe vs. cringepics). The surprising affix is *true*. Although it seems to suggest superiority and separation, explorers actually become more active in the original commu-

nity in this case, too.

Discussion. Our findings resonate with Tan and Lee's (2015) results that users who continually explore new communities are, on average, more active than users who don't. However, no causal relationship can be established that explains this result: exploration does not necessarily *cause* increased activity. However, in our dataset, exploration appears to be a strong signal of interest level.

Variations between explorers with different activity levels

We have established that users tend to be relatively more active after exploration, and have examined the variation across different categories and pairs. However, how does this effect differ for users with different pre-interaction levels? One could imagine that activity level prior to exploration affects whether or not users are more active after exploring. For example, upon discovering an alternative community, it's possible very active users might remain more attached to their home community, whereas relatively inactive users might not have the same level of commitment.

To address this question, we split users into pre-interaction quartile levels within their spinoff pair, so that the users with the least number of pre-interactions are put in bin one, and the most number of pre-interactions are put in bin four. We then compute exploration effect for users in each quartile.[8]

Figure 7 presents the fraction of users who had more post-interactions than pre-interactions respectively for explorers and nonexplorers in several popular subreddit pairs. In general, the relative effects of exploration appear to be different

---

[8]We have previously referred to Equation 2 as exploration effect, but plot p_e and p_{ne} separately in these plots under the same name.

based on how active users are, but there are complex and varied relationships between user activity level and how much defection matters; these relationships differ based on which spinoff pair is considered. Since we split users further into quartiles, the amount of data is not sufficient to reach conclusions for all pairs.

One relatively *consistent* pattern across pairs is that explorers with the highest pre-activity level usually have a smaller difference from the nonexplorers compared to explorers in the lowest quartile, as shown in the left three figures in Figure 7, although this is not true for Figure 7d.

The trend of how the fraction or the difference changes with different pre-activity quartile is even more complex. Consider the case of Figure 7a; this figure illustrates that for users with low activity levels (first/second quartiles) exploring is much more indicative of increased future activity than not exploring, and the difference is much less apparent for users with high activity levels – exploring and not exploring are associated with more similar levels of activity for users in the third/fourth quartiles. An explanation of this result is that exploration matters more to low activity users, and appears to excite them more easily. While exploration still matters to high activity users, the difference is not as large.

Note that other pairs exhibit different patterns. For Science vs AskScience (Figure 7b) and Android vs androidapps (Figure 7c), the most active users (those in the 4th quartile) appear to experience a slight "dip" in terms of the exploration effect.

Related Work

While there has been considerable interest in the topic resulting from the social sciences (e.g., (Hurtado 1997; Berry 1997)), the study of situations wherein users engage with *multiple*, distinct communities represents a relatively new but increasingly relevant research area for computer scientists. Indeed, Kim (2000) argues that a growing Web *needs* subdivisions, while Jones and Rafaeli (2010) also argue that an effective community splitting strategy is essential for virtual communities and online discourse to thrive. Furthermore, Birnholtz et al.'s (2015) study of confession groups on Facebook could be viewed in the context of "place" style affixs. While our study has a different focus, our findings mirror those of Tan and Lee (2015) who found that users who explore a greater number of communities are among the most active across meta communities.

A number of studies have examined multi-community platforms in different contexts. Subcommunity survival (Turner et al. 2005; Iriberri and Leroy 2009; Kraut et al. 2012) is sometimes framed in the context of a meta-community. Also, Fisher et al. (2006) find that different newsgroups exhibit different conversation patterns, though don't examine if the same users behave differently across platforms (as in (Vasilescu, Filkov, and Serebrenik 2013)). Finally, Adamic et al. (2008) examine the quality of user answers across different categories of Yahoo Answers.

Despite exhibiting some undesirable upvoting patterns (Gilbert 2013), Reddit itself has been used as a data source in various contexts. For instance, the study of altruistic requests (Althoff, Danescu-Niculescu-Mizil, and Juraf-

sky 2014), the study of domestic abuse discourse (Schrading et al. 2015), and the work about post titles (Lakkaraju, McAuley, and Leskovec 2013) demonstrate that useful information can be learned from Reddit comments and upvotes.

Conclusion

In this work, we use a dataset of all posts and comments from Reddit over an eight-year period to explore the space of naming affixes that lead to highly related communities on the platform. After building a taxonomy, we examine the early participants and other temporal aspects of the pairs, and introduce the idea of a spinoff community being "born out" of its unaffixed parent. Finally, we present the surprising result that users who explore in spinoff communities generally become relatively *more* active in their home communities instead of being distracted. We also find that the magnitude of this effect (and sometimes its sign) depend on the type of communities and how active a user was prior to exploration.

There are several directions of possible future work. First, it would be interesting to examine more closely the *origins* of highly related communities. If a community is created because of a disagreement (e.g., Zachary's Karate Club (1977)) one could potentially identify general characteristics of a increasing unrest prior to a fission. Also, it would be interesting to delve deeper into differences between discourse on content in highly related communities pairs; how does discussion on TrueAtheism differ from discourse on Atheism, for example. It would be useful for community organizers if we can detect when a spinoff community is necessary or beneficial. Furthermore, it is an important direction to understand the mechanism behind our observation that users who explore in spinoff communities generally become relatively *more* active in their home communities. This could be potentially useful for community organizers to identify complementary communities.

Finally, we note that our consideration has presupposed a pairwise framing, i.e., we always assumed a *pair* of communities. In some cases, we noted more complex phenomena underlying community creation. For example, a number of communities, all made with the affix srs were created in a short period of time. Also, the world of pokemon subreddits may consist of multiple affixes that lead to different subdivisions. In general, one could generalize pairwise interactions to explore more complex relationships between communities.

Appendix: Sampling Method for Control Users

The goal of this section is to describe how we sample a control user u_{ne} corresponding to each each exploring user u_e. Ultimately, to compute the exploration effect, we need find someone who never posts in the new subreddit, but has a

similar amount of actions in the same time period. To choose this similarly active, nonexplorer user, we sample u_{ne} as follows:

1. From the set of all nonexplorer users, find the subset who have an interaction in the original community within 24 hours of u_e's exploration time t. Let these interactions occur at time t'. If a nonexplorer user has more than one interaction between $t - 24$ hours and $t + 24$ hours, take the closest to t.

2. Find the user u_{ne} in this candidate set that minimizes the difference between their own number of pre-interactions (re-centered at their t') and u_e's. Specifically, if we let $p(u, t_a, t_b)$ be the number of interactions of user u in the original subreddit between t_a and t_b we find the loyal user

$$\underset{u_{ne}}{\operatorname{argmin}} |p(u_e, t - w, t) - p(u_{ne}, t' - w, t')|$$

3. If this difference is less than 5% of u_e's pre-interactions, a similarly active user u_{ne} has been successfully sampled.

Figure 5 demonstrates a pair of users that could be plausibly sampled in this manner. Both the light and dark users have the requisite 5 pre-interactions, and the light user makes a post within 24 hours of the dark user's first exploration.

After sampling k such user pairs $\{\langle u_{i,ne}, u_{i,e}\rangle\}|_{i=1}^k$ for a given pair of subreddits[9], we first compute the proportion of exploring/nonexplorer users whose activity increased, i.e., with more post-interactions than pre-interactions. For instance, this fraction for exploring users is computed as

$$p_e := \frac{1}{k} \sum_{i=1}^k \mathbb{1}\left[post(u_{i,e}) > pre(u_{i,e})\right] \qquad (1)$$

Finally, for each spinoff pair of communities, the quantity we are interested in is

$$p_e - p_{ne} \qquad (2)$$

We generally call Equation 2 "exploration effect". A larger exploration effect indicate that an explorer is more active in the *original* subreddit after posting to the splinter subreddit, when compared to a similarly active nonexplorer. In Figure 7, we plot p_e and p_{ne} separately, whereas in Figure 6 we plot $p_e - p_{ne}$.

References

Adamic, L. A.; Zhang, J.; Bakshy, E.; and Ackerman, M. S. 2008. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of WWW*, 665–674.

Althoff, T.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of ICWSM*.

Backstrom, L.; Huttenlocher, D.; Kleinberg, J.; and Lan, X. 2006. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of KDD*, 44–54.

Berry, J. W. 1997. Immigration, acculturation, and adaptation. *Applied psychology* 46(1):5–34.

Birnholtz, J.; Merola, N. A. R.; and Paul, A. 2015. "Is It Weird to Still Be a Virgin": Anonymous, Locally Targeted Questions on Facebook Confession Boards. In *Proceedings of CHI*.

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 *Proceedings of tWWW*, 307–318.

Ducheneaut, N.; Yee, N.; Nickell, E.; and Moore, R. J. 2007. The life and death of online gaming communities: a look at guilds in world of warcraft. In *Proceedings of CHI*, 839–848.

Fisher, D.; Smith, M.; and Welser, H. T. 2006. You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of HICSS*. IEEE.

Gilbert, E. 2013. Widespread underprovision on reddit. In *Proceedings of CSCW*, 803–808.

Hessel, J.; Schofield, A.; Lee, L.; and Mimno, D. 2015. What do vegans do in their spare time? latent interest detection in multi-community networks. *NIPS Workshop on Networks in the Social and Information Sciences*.

Hurtado, A. 1997. Understanding multiple group identities: Inserting women into cultural transformations. *Journal of Social Issues* 53(2):299–327.

Iriberri, A., and Leroy, G. 2009. A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)* 41(2):11.

Jones, Q., and Rafaeli, S. 2010. Time to Split, Virtually: 'Discourse Architecture' and 'Community Building' Create Vibrant Virtual Publics. *Electronic Markets* 10(4):214–223.

Kairam, S. R.; Wang, D. J.; and Leskovec, J. 2012. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of WSDM*, 673–682. ACM.

Kim, A. J. 2000. *Community Building on the Web: Secret Strategies for Successful Online Communities*. Addison-Wesley Longman Publishing Co., Inc., 1st edition.

Kraut, R. E.; Resnick, P.; Kiesler, S.; Burke, M.; Chen, Y.; Kittur, N.; Konstan, J.; Ren, Y.; and Riedl, J. 2012. *Building successful online communities: Evidence-based social design*. Mit Press.

Kumar, R.; Novak, J.; and Tomkins, A. 2010. Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*. Springer. 337–357.

Lakkaraju, H.; McAuley, J. J.; and Leskovec, J. 2013. What's in a name? Understanding the interplay between titles, content, and communities in social media.

Schrading, N.; Alm, C. O.; Ptucha, R.; and Homan, C. 2015. An analysis of domestic abuse discourse on reddit. 2577–2583.

Singer, P.; Flöck, F.; Meinhart, C.; Zeitfogel, E.; and Strohmaier, M. 2014. Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community? In *WWW (Companion)*.

Tan, C., and Lee, L. 2015. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of WWW*, 1056–1066.

Turner, T. C.; Smith, M. A.; Fisher, D.; and Welser, H. T. 2005. Picturing usenet: Mapping computer-mediated collective action. *Journal of Computer-Mediated Communication* 10(4):00–00.

Vasilescu, B.; Filkov, V.; and Serebrenik, A. 2013. StackOverflow and GitHub: Associations between software development and crowdsourced knowledge. In *Proceedings of SocialCom*, 188–195. IEEE.

Zachary, W. W. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33(4):452–473.

---

[9]We discard the pair of communities if $k < 100$.