

# Evolution of Reddit: From the Front Page of the Internet to a Self-referential Community?

Philipp Singer\*  
Graz University of Technology  
philipp.singer@tugraz.at

Fabian Flöck\*  
Karlsruhe Institute of Technology  
floeck@kit.edu

Clemens Meinhart  
Graz University of Technology  
c.meinhart@student.tugraz.at

Elias Zeitfogel  
Graz University of Technology  
elias.zeitfogel@student.tugraz.at

Markus Strohmaier  
GESIS & U. of Koblenz  
markus.strohmaier@gesis.org

## ABSTRACT

In the past few years, Reddit – a community-driven platform for submitting, commenting and rating links and text posts – has grown exponentially, from a small community of users into one of the largest online communities on the Web. To the best of our knowledge, this work represents the most comprehensive longitudinal study of Reddit’s evolution to date, studying both (i) how user submissions have evolved over time and (ii) how the community’s allocation of attention and its perception of submissions have changed over 5 years based on an analysis of almost 60 million submissions. **Our work reveals an ever-increasing diversification of topics accompanied by a simultaneous concentration towards a few selected domains both in terms of posted submissions as well as perception and attention.** By and large, our investigations suggest that Reddit has transformed itself from a dedicated gateway to the Web to an increasingly self-referential community that focuses on and reinforces its own user-generated image- and textual content over external sources.

## Categories and Subject Descriptors

H.3.5 [Information Storage and Retrieval]: Online Information Services

## Keywords

Reddit; frontpage; evolution; growth; longitudinal; online community; attention; perception; discussion; diversification; self-reference

## 1. INTRODUCTION

Since its founding in 2005, Reddit has grown into one of the largest online communities on the web. As of this writing, the site has more than 112 million unique visitors from over 195 countries

\*Both authors contributed equally to this work.

each month.<sup>1</sup> It is ranked by *Alexa.com* as the 69<sup>th</sup> and 27<sup>st</sup> most popular website in the world and the U.S., respectively.<sup>2</sup> On Reddit, users can post links to external websites *or* submit textual content directly hosted on Reddit, so-called self submissions or self posts. **Other “Redditors” – a neologism combining “Reddit” and “editor” – can then up- and downvote the posted items, contributing to an ever-changing ranking of the “hottest” submissions.** Users can comment on every submission as well as create their own sub-communities named “subreddits” – each being independent, dedicated to a specific topic and moderated by volunteers. Equipped with these features, Reddit was intended to capture and rank all kinds of diverse content collected from the Web by promoting the best parts via its voting process. Reddit’s original claim is that the site represents *“the front page of the Internet”* – suggesting that it **acts as a gateway to (the best) content available on the Web.** Today, this declaration is still prominently featured in the HTML title of *Reddit.com*. With this mission statement, the platform has exhibited exponential growth in terms of submissions between 2008 and 2012, as evident in Figure 1.<sup>3</sup> Yet, it remains for the most part unclear whether the initial design intentions behind Reddit are still relevant today, or whether the system has evolved to accommodate other purposes. In the following we aim to address this question.

**Research questions:** Specifically, we address two issues:

(i) *Longitudinal analysis of user submissions:* We examine in detail how user submissions to Reddit have evolved over the course of five years. Regarding diversity of subreddits, top-level domains of posted links and types of content allow us to evaluate whether and how the focus of user posts to Reddit has changed over time. (ii) *Longitudinal analysis of perception and attention:* To gauge whether and how perception and attention by the Reddit community developed, we analyze voting and commenting patterns, enabling us to assess what kind of submissions received attention by Redditors over time. In this work, we use a large-scale dataset containing all submissions to Reddit 2008-2012 (close to 60 mio submissions). A succinct user survey supplements our analysis.

**Contributions & results:** To the best of our knowledge, this work represents the most comprehensive longitudinal study of Reddit’s evolution to date, studying both (i) how user submissions have evolved over time and (ii) how the community’s allocation of atten-

<sup>1</sup><http://www.reddit.com/about/>, as of Feb. 02<sup>th</sup>, 2014

<sup>2</sup><http://www.alexa.com/siteinfo/reddit.com>, as of Feb. 02<sup>th</sup>, 2014

<sup>3</sup>Exponential growth being a better fit than a Gompertz model as well as a logistic model, tested on our data described in Section 2.

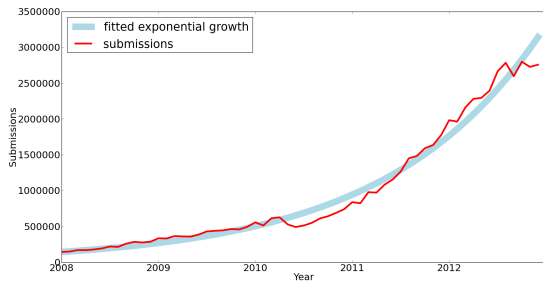


Figure 1: Number of submissions to Reddit each month, ranging from Jan. 2008 to Dec. 2012 (red line). The blue line denotes the best exponential fit for the growth, ongoing until the end of 2012.

tion and its perception of submissions have changed over 5 years based on an analysis of almost 60 million submissions. Our analysis of all Reddit submissions from 2008-2012 reveals an ever-increasing diversification of topics (i.e., subreddits) accompanied by a simultaneous concentration towards a few selected domains and types of submissions (self and image). This suggests that Reddit has transformed itself from a dedicated gateway to Web content to an increasingly diverse, self-referential community that focuses on and reinforces its own user-generated content over external sources. Our work sheds light on formerly unknown dynamics of Reddit and represents an important step towards a deeper understanding of Reddit and similar platforms.

**Structure of this paper:** In the next section, we describe our dataset and methods. We present our results in Sections 3 and 4 and discuss them in Section 5. After listing related work in Section 6, conclusions and future work are summarized in Section 7.

## 2. DATASET AND METHODS

In this section, we introduce our dataset<sup>4</sup> and describe the method used for categorizing the content submitted to Reddit into six categories, before explaining the design of our user survey.

### 2.1 Description of the dataset

We analyze data consisting of all submissions posted to Reddit from January 2008 to December 2012 crawled through Reddit’s API.<sup>5</sup> The metadata of each submission (i.e., title, author, up- and downvotes, number of comments, the link or text it contained and the submission time) were collected around 1-2 months after the initial submission (i.e., when they get blocked from voting) as the metadata has most likely been settled after this period. Overall, we analyze 58,874,22 submissions (14,979,707 self posts) in 125,662 distinct subreddits from 4,910,850 different authors linking to 1,841,239 distinct domains on the Web.

**Categorizing submissions on Reddit:** We also provide a categorization of the content of the links that are submitted for facilitating an analysis of the types of content on Reddit. We manually classified the 100 most frequently submitted domains which represent 69% of all submissions (including self posts), into six categories: *text*, *image*, *video*, *audio* and *misc* and the last category *self*, which accounts for all self posts in the dataset (25% on their own). Domains were assigned a single category after examining their main purpose or type of content. The *text* category covers everything from news-sites to blogs with focus on textual content and even encyclopedias (e.g., Wikipedia). *Image*, *video* and *audio* are mainly hosting services and content providers for specific types of

media; the most used examples in the dataset would be *Imgur.com*, *Youtube.com* and *Soundcloud.com*, respectively. The *misc* category covers domains that do not clearly fit into one of the other categories and comprises, e.g., link shorteners like *Tinyurl.com* or universal hosting services like *Amazon Web Services*.

### 2.2 Description of the user survey

After the main data collection, a short auxiliary user survey with questions regarding certain aspects of this paper was posted to the subreddits *r/theoryofreddit* and *r/samplesize*, as their user communities are very open to providing answers to questionnaires.<sup>6</sup> This particular, limited sampling and the self-selection of respondents must be taken into account when interpreting the results. Our analysis showed, however, no notable difference between the answer patterns of the two subreddits and will henceforth be reported in aggregate. The survey ran from Nov. 24 until Dec. 1, 2013 and yielded 1,004 responses (Note: some questions were optional and not answered by all users). We filtered obvious spam answers from the results, leaving 969 answers: 66% from *r/theoryofreddit* and 34% from *r/samplesize*.<sup>7</sup> Questions and results from the survey will be reported as supplemental information in selected sections.

## 3. DIVERSITY AND SELF REFERENCE

As Reddit has experienced exponential growth (cf. Section 1), it is not unlikely that the internal dynamics of Reddit have evolved correspondingly and thereby affected the character of the site as a whole and particularly its function as the “front page of the Internet”. We are thus interested in investigating the current distribution and change over time of three important aspects of Reddit: (a) subreddits, (b) linked domains and (c) types of linked content.

### 3.1 The diversification of subreddits

As outlined in Section 1, one main aspect of Reddit today is the existence of thousands of distinct – mostly user-created – subcommunities, also called subreddits. We measure the popularity of subreddits by counting committed submissions to a particular subreddit at a specific time (monthly). In Figure 2a the development of all active subreddits (i.e., with at least one submission) is depicted, with their relative size in percent compared to the overall size in total submissions on Reddit at a specific time. We only visualize the 20 largest subreddits with distinct colors and combine the rest in brown color. We can observe that a fragmentation of submissions into an ever-increasing amount of distinct subreddits has taken place since Reddit’s inception. In the last month of our data set in 2012, 32,202 subreddits received one or more submissions while only 213 did so at the beginning of 2008. The 20 biggest subreddits at the end of 2012 contained less than 40% of all submissions to Reddit, while they contained around 70% and 80% in mid-2010 and mid-2008, respectively. Measured as a Gini coefficient, the concentration of submissions over subreddits decreased slightly from 0.97 in mid-2008, over 0.95 in mid-2010 to 0.94 at the end of 2012. These findings point, at first glance, to a strong diversification of topics represented by the different subreddits, whose establishment does in fact articulate the more explicit need of a part of the user base for certain dedicated thematic spaces. However, many topics and discourses might have existed previously as part of one of the broader themed subreddits, especially *r/Reddit.com*, which served as the default posting space in the early phase of Red-

<sup>4</sup>Dataset access can be requested at <http://www.philippsinger.info/reddit/>.

<sup>5</sup><http://www.reddit.com/dev/api>

<sup>6</sup>The full questions plus additional information can be found at <http://people.aifb.kit.edu/ffl/redditsurvey>

<sup>7</sup>Unreasonable values (like “25 hours per day”) and text (nonsensical, flaming) were used as indicators of spam.

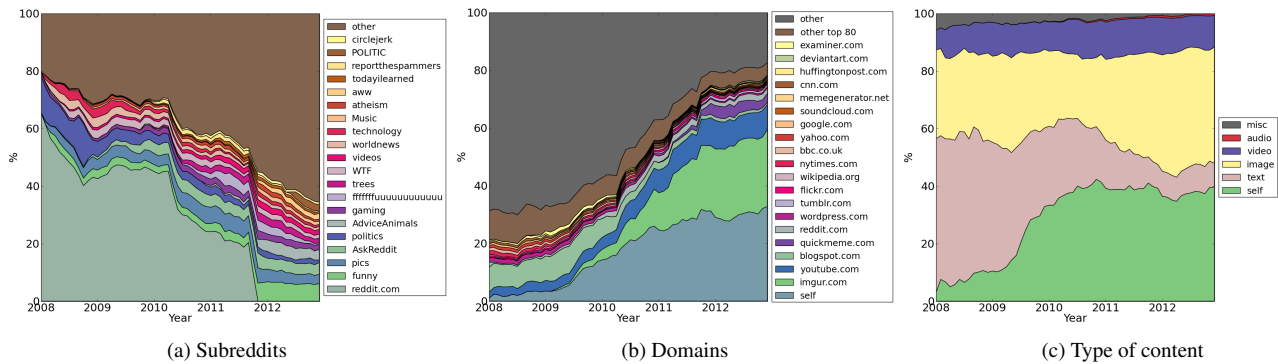


Figure 2: Evolution of submissions per month over subreddits, domains and types of content, 2008-2012.

dit. Figure 2a unveils *r/Reddit.com*'s gradual demise. When user-founded subreddits were first introduced at the beginning of 2008, a great quantity of submissions was still committed to *r/Reddit.com*.<sup>1</sup> With more subreddits founded, *r/Reddit.com* kept shrinking, until in October 2011 a set of 20 default subreddits was introduced, which led to *r/Reddit.com*'s deliberate shutdown by the operators.<sup>8</sup>

Overall, an at least *structural diversification* can be noted in the form of a vast increase in subreddits; alas, it cannot be stated with certainty that the general *thematic diversity of submissions* has in fact increased. Some topics might have just been outsourced from more general subreddits similar themed subreddits to sharpen their profile and/or not to get lost in the increasing flood of posts in *r/Reddit.com*. What can be affirmed, however, is that clearly distinct communities around topics had a chance to form in the secluded spaces of the subreddits, each with their own, clear-cut rules and "submission ethics" – visible at the right-hand sidebar of most subreddits. In sum, the subreddit diversification fits sufficiently well with the claim of Reddit representing (the best) content from all over the Web, as the ladder grew exponentially in its already vast content heterogeneity since 2008 and Reddit has seemingly be able to mirror this diversity and built sub-communities around it.

### 3.2 Towards self-reference in submissions and their linked domains

On Reddit users can submit content (a) through self (text) submissions that are stored on Reddit itself and (b) through links to external Web content. The majority of submissions has traditionally been links to external content, as the initial idea of the platform was solely to share links and vote on them. For Reddit to be in fact a "frontpage" for – or a gateway to – content from all over the Web, one would expect an increasing diversity among the link targets (i.e., top level domains) of the submissions similar to what we have seen in the subreddit distribution. Figure 2b reveals that the relative proportion of self submissions has been growing tremendously over time, with an initial boost in mid-2009. A small increase in backlinks to reference older submissions on *Reddit.com* is also visible until mid-2010 and stabilizes at around 2% afterwards. Apparently, the biggest external beneficiary of the content expansion is one image hoster called *Imgur.com* (0% in mid-2008, 7.32% in mid-2010 and 26.6% at the end of 2012) which owes its uprising to the large influx of Reddit posts.<sup>9</sup> *Imgur* was in fact created for the expressed purpose of serving as an image hosting service for Reddit; *Imgur* founder Alan Schaaf stated in 2009 that he designed the platform

<sup>8</sup><http://blog.reddit.com/2011/09/independence.html>

<sup>9</sup><http://www.mediaite.com/online/imgur-accounts-alan-schaaf-interview/>

as he was not satisfied with other available image hosts.<sup>10</sup> Since its inception, *Imgur* is has not only risen to be the primary image host for Reddit, it has actually become a central aspect of Reddit's culture; so much in fact, that it is close to being the sole image hosting option for Reddit posts that is accepted by many subreddits, as community discussions reveal.<sup>11</sup>

The external domain exhibiting the second largest expansion in linked content is *Youtube.com* (2.37% in mid-2008, 6.24% in mid-2010 and 8.68% at the end of 2012), followed by *Quickmeme.com* (0% in mid-2008, 0% in mid-2010 and 3.05% at the end of 2012), a website providing a an easy-to-use service for creating meme images with custom personal messages. The two domains that suffered most were *Blogspot.com* (7.68% in mid-2008, 3.03% in mid-2010 and 0.83% at the end of 2012) and *Wordpress.com* (1.41% in mid-2008, 0.99% in mid-2010 and 0.49% at the end of 2012), both hosting mainly blogs, i.e., text based content.

Thus, while Section 3.1 unveiled that the content on Reddit frays out into more and more subreddits – i.e., thematic subspaces – over time, we now see that submissions to external content concentrated more and more to just a few domains, mainly self and *Imgur.com*. The Gini coefficient for concentration accordingly increased notably from 0.78 in mid-2008, over 0.83 in mid-2010 to 0.95 at the end of 2012 – computed over URLs that received submissions that month. The overall diversity of linked domains has meanwhile been keeping up fairly in accordance with the submission growth, with the total number of distinct domains being 34,082 in mid-2008, 68,577 in mid-2010 and 103,660 at the end of 2012. The shifting focus on self-referential posts thus evolved parallel to an otherwise still diverse spectrum of linked domains.

### 3.3 A shift to "self" and images

To better understand the distribution of links over external domains, we analyze the type of content these URLs usually host and how the popularity of each content type has progressed. To this end, we make use of the classifications of top-level domains provided in Section 2.1 into image, video, text, audio, misc and self.

The progression over time – represented via the relative proportion of each category, cf. Figure 2c – confirms that self posts have not always been the favorite kind of submission. From 2008 to mid-2009 the majority of submissions were linking to external textual content. Over time, the (likewise textual) self submissions exceeded the number of external textual submissions, consistent with the decline of *Blogspot.com* and *Wordpress.com*. Yet, while some material from blog sites or even news portals formerly linked might

<sup>10</sup><http://www.reddit.com/tb/7zlyd/>

<sup>11</sup>Cf. discussion threads on [http://www.reddit.com/r/TheoryOfReddit/search?q=imgur&restrict\\_sr=on](http://www.reddit.com/r/TheoryOfReddit/search?q=imgur&restrict_sr=on)







the early phase of Reddit, visitors are nowadays much more likely to end up consuming self-referential content instead of finding their way through the proverbial gateway back out into the Web. This is on the one hand because of the probability of said content to make up large parts of ranked submission lists, due to its sheer volume. But it can also be attributed to users' own affinity for such content, which they seek out or are at least prone to vote on and thereby catapulting it up the "hot" lists of Reddit, instigating a self-reinforcing cycle. It is also valid to assume that with more exposure to community-centric content, users become more involved in Reddit's "biotope", producing such content themselves in turn. The centrality of Reddit for the information diet of its users is underlined by our user study (Table 2), exemplifying that Reddit is (a) the main website for certain topics for many users, (b) is often visited daily and (c) is rather used with no specific information need in mind, leaving users susceptible to the suggestions by the system.

Arguably, the community aspect of Reddit is becoming more important and images and self posts are the prime communication means between its members, an assumption strengthened by the survey results in Table 1, highlighting messaging, entertainment and pictures, while still prominently mentioning news and the portal function. We can witness the growth and increasing self-definition of a community – a phenomenon we could only briefly revisit in this paper. Without judging this evolution in one way or another, it certainly changes the nature of Reddit as a link-sharing platform, affecting the once straightforward and simple link exchange in ways that merit further study.

## 6. RELATED WORK

Lakkaraju et al. [2] studied how titles, submission times and community choices of image submissions affect the success of the content by investigating resubmitted images on Reddit, showing that good content can speak for itself, although a good title has a positive effect on popularity. Gilbert [1] investigated resubmissions of content to Reddit and compared their eventual voting score, finding that identical links are ignored by the community several times before achieving popularity. Weninger et al. [6] focus on comment threads on Reddit, showing that highest scoring comments are mostly submitted at early stages of the discussion. For the similar platform *digg.com*, studies comparable to the above have been conducted, some juxtaposing Digg and Reddit in specific aspects (e.g., Lerman [3]). In addition, the Reddit community itself has shown an interest in the evolution of the platform. An example can be found in a blog post [5] that has looked at the evolution of submissions via subreddits (cf. Figure 2a), suggesting a trend towards diversification. While this blog post presents interesting initial insights into Reddit's development, our work advances them by systematically studying and comparing the evolution of domains, content types, the perception of submissions via scores, comments and votes and other aspects coupled with a user survey (969 respondents) in order to get a more comprehensive understanding of Reddit's evolution.

Table 2: User study results 2. Left: Percent of participants naming Reddit as their main source of a specific content (multiple choice). Right: Mean answer values for various questions. Cf. fn. 6

Main website for content?	%	Question	Mean
Entertainment/Distracton	90	# of websites visited	11.80
Education, advice, learning	61	daily	
News	59	Rank of Reddit among	1.98
Social interaction, discussion	46	top 10 daily sites	
File sharing	5	Likert 1-7, 1= "Looking	5.27
Not main site for any content	6	for smth. specific", 7=	
Other	5	"Just exploring"	

## 7. CONCLUSIONS

To the best of our knowledge, this work represents the most comprehensive longitudinal study of Reddit's evolution to date, studying both (i) how user submissions have evolved over time and (ii) how the community's allocation of attention and its perception of submissions have changed over 5 years based on an analysis of almost 60 million submissions. Our main findings are threefold: (i) *Increasing diversification of topics*: we found that Reddit has evolved from a small community capturing a broad topic area to a platform covering a large number of distinct sub-communities with specialized interests and topics. (ii) *Concentration towards a few domains*: we can observe that over time, submissions and attention (comments) increasingly focused on two domains, i.e., *Imgur.com* and self, i.e. the Reddit community increasingly reinforces its own user-generated image- and textual content over external sources. These results suggest that (iii) *Reddit has transformed from a dedicated gateway to the Web ("The front page of the Internet") to an increasingly self-referential community*. Our results are both reflected by how submissions are posted on Reddit as well as how Redditors perceive the ever-changing arrangement of submissions and how they divide their attention among them. From our analysis it remains unclear whether the observed changes in Reddit's community are the result of a conscious effort (e.g., by the operators of the platform or by influential subreddits), or whether the community merely gradually drifted towards a more self-referential mode of operation. We leave answering this question to future work.

Overall, our work shows how an online community with (a) high degrees of freedom for users (e.g., voting, commenting or creating sub-communities) and (b) exceptional growth over several years may dramatically change its nature and focus over time. While we delivered a preliminary analysis of Reddit as an example of a large and growing community, we hope that our work inspires others to expand this line of research to more in-depth studies of Reddit and other comparable community platforms (such as *hackernews*<sup>13</sup>).

**Acknowledgments** We thank Jason Baumgartner of *Redditanalytics.com* for supplying us with the Reddit log data. This work was partially funded by the FWF Austrian Science Fund Grant I677.

## 8. REFERENCES

- [1] E. Gilbert. Widespread underprovision on reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work, CSCW '13*, pages 803–808, New York, NY, USA, 2013. ACM.
- [2] H. Lakkaraju, J. McAuley, and J. Leskovec. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.
- [3] K. Lerman. Social networks and social information filtering on digg. In *Proceedings of International Conference on Weblogs and Social Media*, 2007.
- [4] B. Nonnecke and J. Preece. Lurker demographics: Counting the silent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '00*, pages 73–80, New York, NY, USA, 2000. ACM.
- [5] R. Olson. Retracing the evolution of Reddit through post data, <http://dx.doi.org/10.6084/m9.figshare.650851>, Mar. 2013.
- [6] T. Weninger, X. A. Zhu, and J. Han. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Proceedings of the 2013 IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM 2013)*, Aug. 2013.

<sup>13</sup><https://news.ycombinator.com/>