Building a kinder super highway: Online group behavior driven by platform design and social policy

A Dissertation Presented

by

Milo Zappa Trujillo

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Complex Systems and Data Science

August, 2024

Defense Date: July 18th, 2024
Dissertation Examination Committee:

James P. Bagrow, Ph.D., Advisor
Laurent Hébert-Dufresne, Ph.D., Advisor
Alexis Brieant, Ph.D, Chairperson
Jeremiah Onaolapo, Ph.D.
Holger Hoock, DPhil, Dean of the Graduate College

Abstract

Much of human socialization occurs online, and is mediated by telecommunications platforms, particularly social media. These platforms both facilitate and restrict interaction in two ways: first, through the technical affordances they offer, such as conversation trees or direct messages or community self-moderation and voting; and second, through social policy, particularly regarding what content is permissible on a platform and how infractions are penalized. My work engages with platform influence over group social behavior through a series of case studies and through introducing purpose-built methodology.

I begin by examining the influence GitHub exhibits over open-source software development by contrasting the development practices of projects hosted on and off of the platform, showing how increased project discoverability and lower barriers to participation increase "drive-by" contributions from non-project-members, yet GitHub projects tend to have fewer active team members and shorter maintenance lifespans than their off-platform peers.

Next I study the impact of Reddit's content policies regarding hate speech and harassment. My team examined behavioral changes after Reddit banned thousands of communities, illustrating how top power users and the broader community population change their activity and in-group vocabulary usage after such interventions. The heterogeneous results suggest that community-level bans are effective at disrupting only some kinds of communities, and are ineffective at curbing other hostile behavior.

I pivot from platform influence on groups to how groups influence one another by introducing a metric for measuring group-level social centralization. This metric identifies how far a platform tends towards an oligarchy where the largest communities are well-integrated with the platform and are involved with most users. This incorporates both the distribution of community sizes and their insularity. I describe a cumulative "disruption" metric, which removes communities largest to smallest and measures the impact on the remaining population. I demonstrate this metric on five real-world social and collaboration networks, and a variety of synthetic networks, showing how it distinguishes between different kinds of platforms.

Online social platforms exist in an ecosystem, where users and communities can migrate between sites and technologies. Therefore, the affordances offered by and social policies instated on one platform can impact behavior on other platforms. In my final chapter, I propose a group linguistic fingerprint approach to identifying communities even as they migrate between platforms. Such a fingerprint would face a number of challenges, and this chapter is concerned with distinguishing between the vocabulary of group members and the vocabulary of people discussing a group.

Citations

Material from this dissertation has been published in the following form:

Trujillo, M.Z., Hébert-Dufresne, L., Bagrow, J.. The penumbra of open source: projects outside of centralized platforms are longer maintained, more academic and more collaborative. EPJ Data Science 11.1 (2022) 31

Trujillo, M.Z., Rosenblatt, S.F., de Anda Jáuregui, G., Moog, E., Paul V Samson, B., Hébert-Dufresne, L., Roth, A.M.. When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban. WOAH 2021 (2021) 164

Trujillo, M.Z., Minot, J., Rosenblatt, S.F., de Anda Jáuregui, G., Moog, E., Roth, A.M., Paul V Samson, B., Hébert-Dufresne, L.. Distinguishing In-Groups and Onlookers by Language Use. Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis. 2022 157-171

Material from this dissertation has been submitted for publication to Online Social Networks and Media on April 24$^{\text{th}}$, 2024:

Trujillo, M.Z., Hébert-Dufresne, L., Bagrow J.. Measuring Centralization of Online Platforms Through Size and Interconnection of Communities

Turns out the real PhD thesis was in fact Not the friends I made along the way. Not even close, apparently. Turns out I need to write ideas down?? That's pretty cringe, I just did all the thinking via thoughts in [my] head and I trust them.

-Doctor Tom de Prinse, *Explosions&Fire*

Acknowledgements

First, I wish to thank my many mentors: Laurent Hébert-Dufresne, Jim Bagrow, Sibel Adalı, Jean-Gabriel Young, and John Meluso. Their guidance has shaped not only my methods, but the questions I know to ask, and my life is richer for their tutelage.

I further want to share my appreciation for my peers in the Laboratory for Structure and Dynamics and the Complex Data Laboratory, the broader complex systems center, and my colleagues in the Media Landscape and Reddit Bandits labs, both for our research collaborations and for creating a culture of mutual support and playful creativity. Among my colleagues, I particularly wish to share my gratitude for Mariah Boudreau, Bryn Loftness, Sam Rosenblatt, Nicholas Roberts, and Jonathan St-Onge, for their immense support and kindness throughout my years here.

For moral and daily support, I thank my family: Jocelyn Chapman and Darryl Trujillo, Oliver Shuey, Wendy Tully-Gustafson, and Erin Solomon. I would not be here without each of you.

I owe great thanks to Google Open-Source through the OCEAN project, and the University of Vermont's department of Mathematics, for funding my research.

Lastly, thank you to Alexandra Elbakyan, whose contribution to free knowledge through the Sci-Hub project has enabled countless scholars' work, including my own.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overview

Whenever we post on a forum, contribute to an open source project, or join a Discord server, we are entering an intentionally structured online community. Our interactions with other people in that space are shaped by the technology of the platform, which determines what user interactions are possible, whether communication occurs in a flat stream of chat messages, a tree of replies and sub-replies, direct messages between users, or subscription feeds of messages. However, our interactions are also determined by social policy: what kinds of interactions are permitted within the community? Who makes in-the-moment moderation and administrative decisions, broader policy decisions, and future goals of the community? How is group-feedback incorporated into that process? A better understanding of both the social impacts of technical choices, and of governance choices, will provide insight into information-flow and implicit collective decision-making within online groups. My work has examined the impact of social policy and platform design on how online non-commercial groups

behave in terms of response to moderation, sustainability, distribution of labor and social capital, and the influence sub-groups have on one another.

I frame my understanding of platform rules through a taxonomy established by Elinor Ostrom's Institutional Analysis and Development Workshop [48]. This taxonomy describes platform "rules" as belonging to one of three categories:

1. **Operational Rules.** These consist of the actions that users can take on a platform, such as posting, commenting, direct messaging, and voting on or reporting content posted by other users. To borrow vocabulary from affordance theory [55] we can call these the actions a platform *affords to* its users. These operations are defined wholly by the platform operators.

2. **Collective Rules.** These consist of the context in which users interact with one another. For example, on many social media platforms there is a "content feed," where posts are promoted based on their age and popularity. Content from one user effectively competes against other content for attention. These rules are outlined by the platform operators, who define how such a content feed algorithm functions, but are driven by the behavior of users on the platform.

3. **Constitutional Rules.** These describe the process by which operational, collective, and constitutional rules may change. On corporate-owned social media this is typically unilateral: management at the company proposes a change, it is implemented by employee engineers, and it impacts users. However, some online communities have a culture of "forking," wherein users can dissent by building a duplicate platform with differing rules, and depending on the present operational rules may be able to bring the contents and social relationships from

the original platform to the fork [143].

When I discuss online communities I am referring to a group of people with shared customs, and typically shared beliefs and interests. These communities occupy one or more social media platforms, but are not defined by them; for example, when the `/r/the_donald` subreddit was banned its members built and migrated to a Reddit-like website, `thedonald.win` [139].

While communities can relocate between, or concurrently occupy several, platforms, it is still useful to describe the partitions on a platform and how the platforms affordances can enable, enforce, or inhibit particular behavior. In a constructive example, Reddit has site-wide social policies enforced by its administrators, and per-community policies enforced by volunteer moderators within each community. This pre-supposes a social hierarchy of users, moderators, and administrators, with increasing governance authority. However, creating more democratic governance structures is often challenging [174]: on many platforms, including Wikipedia and Slack, communities have struggled to build collective decision-making structures on top of software designed for a boolean permissions hierarchy where any user either does or does not have authority to take an action. Therefore, while communities and their presence on a platform should not be conflated, they are inextricably connected, and it is appropriate to study them jointly as a socio-technical unit.

Over the past five years I have studied the impact of content moderation, both by examining inter-platform information-spreading dynamics, and through observing group adaptation in response to moderator interventions. My initial work in this area centered on examining popular content and user interaction behavior on BitChute, an alt-right YouTube clone where banned content creators often migrate

[158, 161]. We tracked how the platform was used to bypass YouTube's policies on election misinformation to spread violating videos on Twitter [31]. By contrast to minimally-moderated alt-tech platforms, Reddit has banned many communities for violating policies on hateful conduct. In chapter 3 we investigate user response to community bans by measuring changes in their activity level and in-group vocabulary usage, among regular and "power" users across fifteen prominent banned subreddits. This line of inquiry can show under what conditions deplatforming is effective at changing user and group behavior, informing platform moderators' and administrators' policies to inhibit hatespeech and radicalization. We followed this study by establishing methodology for distinguishing members of a community from onlookers discussing a community via contextual language markers (see chapter 5). This project is a stepping stone to observing changes in group behavior and structure during cross-platform migrations, where more explicit group membership markers are unavailable.

In chapter 2, we gather a data set of what we believe to be the majority of public repositories across all Git servers outside of GitHub and GitLab, to contrast development patterns on and off the dominant two websites where the majority of open-source software development occurs. By examining trends in commit histories, we find that projects from outside the centralized development platforms had distinct differences: they tend to have more collaborators, are maintained for longer periods, and tend to be more focused on academic and scientific problems. This indicates that development on GitHub is not representative of all of open source, and provides early evidence as to what projects may thrive under different development conditions.

My most recent work examines group-level influence, where on some online plat-

forms the largest groups have disproportionate influence over information flow, while other platforms have less dependence on their largest sub-communities. Lacking an appropriate tool to measure group social influence, we develop a new metric for inferring the impact of removing communities in bipartite networks in chapter 4, which combines the size of communities and their topological role in the rest of the graph. We apply this metric in a five-platform comparison of centralization, demonstrating how community size distributions can mislead researchers into focusing on insular sub-groups that do not represent the broader population.

This dissertation will discuss online group behavior at increasing scales. In chapter 2 I begin with an analysis of how users of GitHub are impacted by operational rules of that platform, including both technical and social affordances. Next, I examine how communities are impacted by platform social policies in chapter 3, through the case study of subreddit deplatforming. In chapter 4, I discuss how communities influence one another, introducing a metric for inter-group centralization on a platform. Finally, I move beyond individual platforms in chapter 5, discussing platform migration and shared community identifiers.

Background

My subjects of interest are social and qualitative: people, their interactions and collective behavior, institutions, and their governance structure and social policies. However, my methods for examining these subjects are distinctly quantitative: web scraping and data mining, statistics and machine learning, natural language processing, and network science. The following chapters assume some familiarity not only

with these domains in their broad strokes, but with specific sub-topics including identifying prominent words by comparing text corpora, or modeling social interactions as weighted bipartite graphs and building null models of similar graphs. Therefore, the rest of this introduction will provide some background in natural language processing, machine learning, and network science, to make those chapters more approachable to readers with a dissimilar background to my own.

1.1 Natural Language Processing

1.1.1 Bag of Words Models

Many natural language processing (NLP) techniques ignore word context, position, and punctuation, focusing exclusively on the frequency of word occurrence. Techniques in this class are called *bag of words* models, since they retain only a collection of words. While highly reductive, bag of words models are mathematically straightforward, and are often used as a precursor to more advanced techniques. For simple tasks, like plotting a change in lexicon over time, bag of words models may be sufficient without invoking more complex machine-learning or word-embedding methods.

1.1.2 Singularization and Lemmatization

English text includes many word variations, such as singular and plural nouns, or verb conjugations. In order to count the number of times a subject is referenced, natural language researchers often seek to combine variations on the same word into a single count. Typical approaches to this problem include removing all punctuation,

standardizing case (lower- or upper-casing all words), reducing pluralized words to their singular format (by removing trailing 's' characters and using a table of irregular plural words), and replacing conjugated verbs with their unconjugated equivalents (i.e. reducing "ran," "run," and "running" to the same word).

1.1.3 Removal of Stop-Words

Some elements of English speech convey structural meaning, but no semantic meaning. For example, "the," "a," "an," and so on. These words are useful when parsing the meaning of a sentence, but are irrelevant in bag of words models, where the number of occurrences of "the" does not tell us much about a body of text. Since these words are among the most common English terms, they add significant noise to any analysis of word frequency. Therefore, most researchers utilize a list of known "stop words," which they discard while counting word occurrences.

1.1.4 Word Frequency Comparison

In my research I often compare the relative prominence of words between two corpora of text to examine differences in vocabulary and subject-matter. If both text samples are of the same length, and contain the same set of unique words, then this comparison is trivial: for each distinct word, calculate the difference in number of occurrences. Words that occur much more often in one corpus than another are notable. However, if corpora are of significantly different sizes, we must compare the frequencies of word occurrence rather than count. This adds an extra challenge when examining words that only occur in one corpora, and therefore appear an infinite

percent more in one text than in the other. Some researchers simply drop words unless they occur in both corpora, but this obscures the growth of new lexicons. An alternative solution is adding each "missing" word to the opposite corpus once; this allows comparison without asymptotic behavior, at the cost of introducing some error to frequency changes, especially for terms that appear infrequently in either corpus. I avoid dropping words or erroneously adding words unless necessary for a comparison to prior work. Instead, I rely on two comparison approaches that handle missing terms in what I consider to be a more principled way: Jensen-Shannon Divergence, and Rank Turbulence Divergence.

Jensen-Shannon Divergence

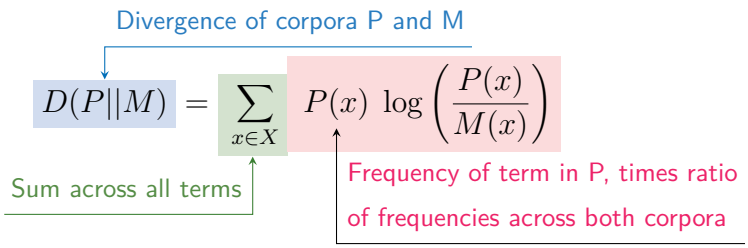
One strategy that allows for non-overlapping lexicon without introducing additional noise is *Jensen-Shannon Divergence (JSD)* [102]. This is an information-theoretic measurement for changes in event frequency, and is commonly used for comparing two probability distributions. In the context of natural language processing we compare the word frequency in each corpus to a combined "mixture" corpus, where word frequencies for each corpus are added and re-normalized. Because the mixture corpus contains the union of all distinct words from each corpora, there is never any asymptotic behavior, and since all frequencies in the combined corpus are derived from word prominence in each individual corpus, there is no added noise. Mathematically, JSD can be described as:

$$JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M) \tag{1.1}$$

Divergence of corpora P and Q

Divergence of P, and Q, from mixture corpus M

Above, M is $\frac{1}{2}(P+Q)$, or the mean frequency of each term across both corpora. D is the *Kullback-Leibler divergence [36],* which is in-turn defined as:



(1.2)

Here, the log of the ratio of frequencies yields an entropic measurement of how much the frequency has changed between the two texts, and multiplying by $P(x)$ means that changes in terms that appear only infrequently matter little, while changes in terms that occur often matters more.

In my work, I am less interested in how far two corpora diverge overall than I am in what the most diverging terms are; that is, words that appear much more prominently in one text than another. In this scenario it is not necessary to calculate the sum of of Kullback-Leibler divergence across all terms. Instead we can sort terms in each corpus by their divergence from the merged corpus M, yielding the terms that most define P and Q. This is the approach used for calculating in-group vocabulary in chapter 3.

With very large probability distributions, such as the word frequencies for all terms used across a subreddit, most individual words appear with vanishing small frequency. In these scenarios it is challenging to interpret an information-theoretic probability divergence score like JSD. This is especially true when the two corpora have a size mismatch, meaning that all terms in the smaller corpus appear with more frequency

than terms in the larger corpus. Additionally, JSD offers no way to "tune" the importance of prominent terms; we expect some common words to overshadow almost all other words, and even small differences in top word usage between corpora can lead to frequency divergences much larger than for almost any other word. However, JSD is well-established and widely used in the scientific community, and can be readily found in many statistical and natural language software packages. For these reasons, we chose to use JSD in chapter 3.

Rank-Turbulence Divergence

Instead of understanding word usage through word frequency, we can alternatively understand a corpus by examining word *rank*. By rank, we mean that the most frequently used word has rank one, the second-most rank two, and so on. Zipf observed that word frequency in English text scales with rank according to a stable relationship [180]:

$$\text{word frequency} \propto \frac{1}{\text{word rank}}$$

In other words, the most common (rank 1) word appears approximately twice as frequently as the second most common word, and three times as frequently as the third most common word.

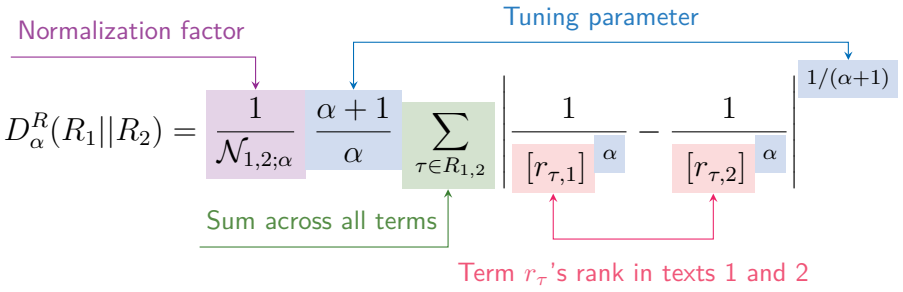
This relationship can be generalized to many other complex systems as:

$$s_r \propto r^{-\varsigma}$$

Where the size of the r^{th} largest component scales according to a decaying power

law with exponent $\varsigma > 0$. Zipf's law is then the particular case where $\varsigma = 1$. However, this generalization is a tangent to my work, because in the following chapters I am specifically interested in comparing word frequency distributions.

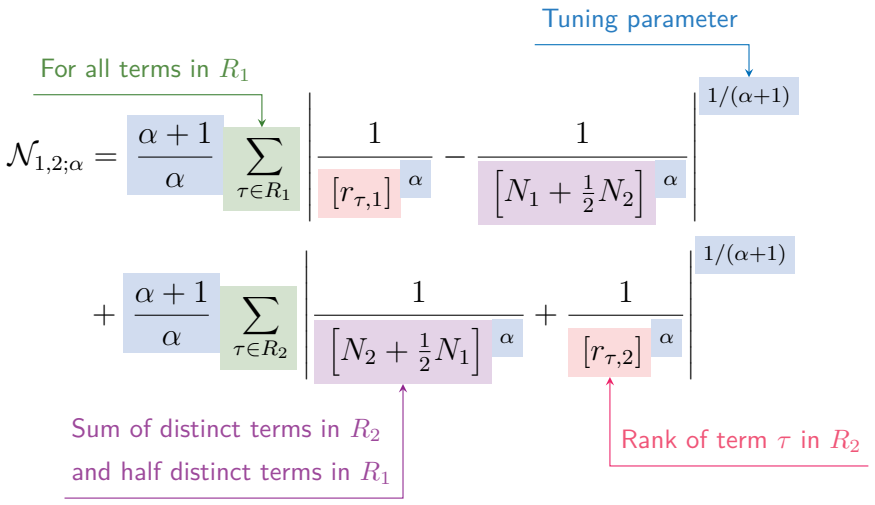
Following from Zipf's law, Rank-Turbulence Divergence (RTD) [41] measures the change in word rank between two corpora (or rather, the reciprocals of ranks, so that low ranks have higher numeric value). It can be written as:



(1.3)

Above, $r_{\tau,s}$ is the rank of element τ (n-grams in our case) in corpora s. The tuning parameter $\alpha \in (0, \infty)$ adjusts the importance of starting ranks: as $\alpha \to 0$, the change in rank of common words and rare words becomes equally important. By contrast, as $\alpha \to \infty$, change in rank of common words becomes more and more important than turbulence in rare word ranks.

For all terms that appear in one corpus, but not the other, Rank Turbulence Divergence adds them to the opposite corpus tied for last-rank. This is necessary so that $1/r_\tau$ is never undefined. However, adding missing terms to each corpus requires re-normalizing the divergence metric by $1/\mathcal{N}$ to guarantee $D \in [0, 1]$, where $\mathcal{N}$ is defined as:



(1.4)

Above, N_1 and N_2 represent the number of unique terms in R_1, and R_2, respectively, and therefore the rank that all missing terms from the other corpus will be assigned.

As with Jensen-Shannon Divergence, my interest is primarily in the most divergent terms between two corpora, and not the overall divergence of two texts. For this purpose, normalization is unnecessary, and Rank Turbulence Divergence simplifies to evaluating the following for all terms in each corpus:

$$\frac{1}{[r_{\tau,1}]^{\alpha}} - \frac{1}{[r_{\tau,2}]^{\alpha}} \tag{1.5}$$

In this expression, the sign indicates whether the word has higher rank in corpus 1 (positive) or corpus 2 (negative), and the magnitude indicates how large a divergence. The divergence score's range depends on both the corpora measured and the α value chosen, and so interpretability is limited without adding normalization. However, ordering terms by their divergence magnitude is sufficient for identifying the terms that are most prominent in each text.

Rank Turbulence Divergence offers interpretability and tune-ability advantages

over Jensen-Shannon Divergence in very large corpora, especially when the corpora have significant size imbalances. For this reason, we use RTD in chapter 5 when comparing the language used in subreddits at different snapshots in time. Regrettably, RTD is presently a more obscure metric than JSD, without as widespread adoption.

1.2 Machine Learning

My research makes occasional use of machine learning, especially classifiers. Classifiers are statistical tools that take a list of numeric *features,* or details about a data point, and produce a categorical *label.* In this section I will introduce a few basic classifiers and means of judging their performance.

1.2.1 Supervised Classifiers

In this dissertation I work exclusively with supervised machine learning binary classifiers, wherein there are only two pre-established categories and the correct labels known for each data point. These classifiers go through a "training" stage, where they are calibrated based on data points with provided labels, and then a "testing" stage, where they are used to predict labels for data points excluded from the training stage. The number of correct and incorrect labels during the testing stage is used to estimate the classifier's efficacy.

In my work, supervised classifiers serve two purposes:

1. Demonstrate that a set of categories are statistically distinct and can be readily distinguished

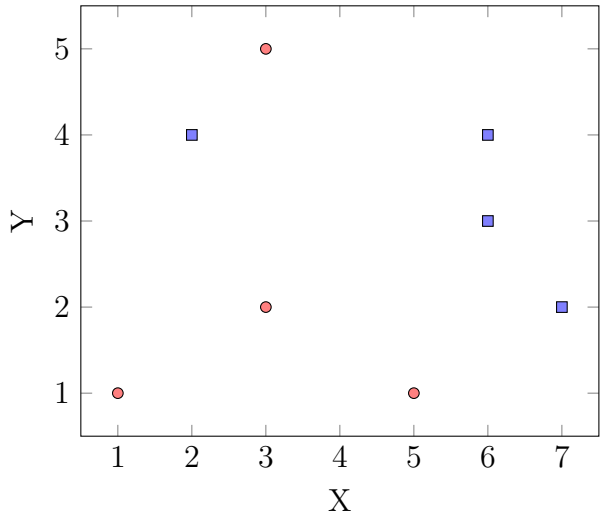
2. Identify which features are the most useful for distinguishing between categories, or in other words, what features most uniquely identify a category

There are many nuances in how to choose a classifier, in how data can be split between training and testing stages to compensate for underrepresented categories, how to prevent "overfitting" on exact observed data points, and how to measure a variety of aspects of classifier strengths and weaknesses. This section will not attempt to address these topics in-depth, but will provide a crash-course in several classifiers used in later chapters and how their performance is measured.

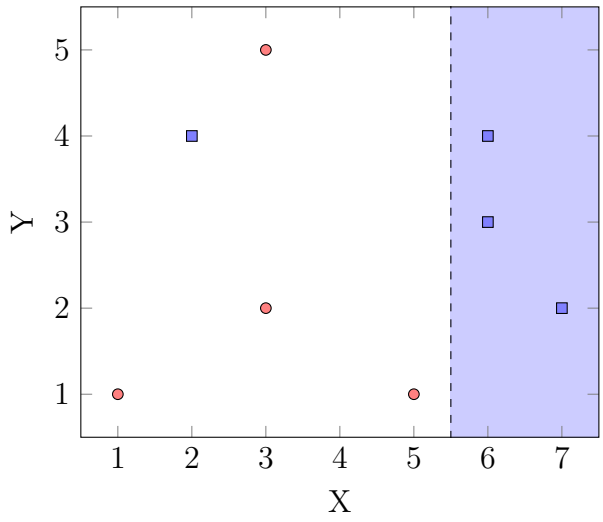
Classification Trees

A particularly intuitive and interpretible classifier is a *decision tree,* more specifically called a *classification tree* when used for classifying data points with discrete labels or a *regression tree* when used to predict continuous values.

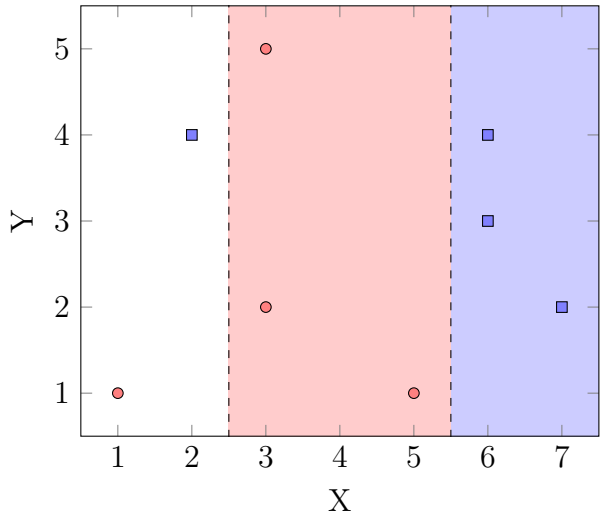
A decision tree treats n features as an $n-$dimensional space, where training data entries are represented as points within that space. For each dimension, the tree algorithm finds an optimal cut-off point that best bisects training data so that points with different labels are separated as much as possible. The best cut across all dimensions is selected. For example, consider the following two-dimensional training data of red circles and blue squares, where we are training the tree to distinguish between the two categories:



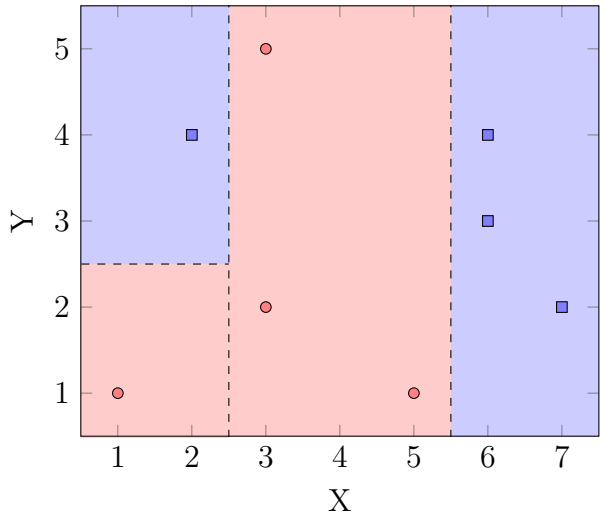
The classification tree finds the best cut across the X- and Y-axes to separate the two classes, as follows:



The right partition contains only the blue square class, but the left partition can be further sub-divided:



Finally, the left-most partition can be subdivided again:



We can visually represent the tree as a flow-chart as shown in fig. 1.1. Now that a classification tree has been "trained" (calibrated or "fit" to the training data), we can present it with new test data points, and the tree will predict whether the test data represents red circles or blue squares by following the trained flow-chart.

Once a tree has been trained, we can retroactively analyze the tree to identify which features best separated the data. Features referenced often near the top of the

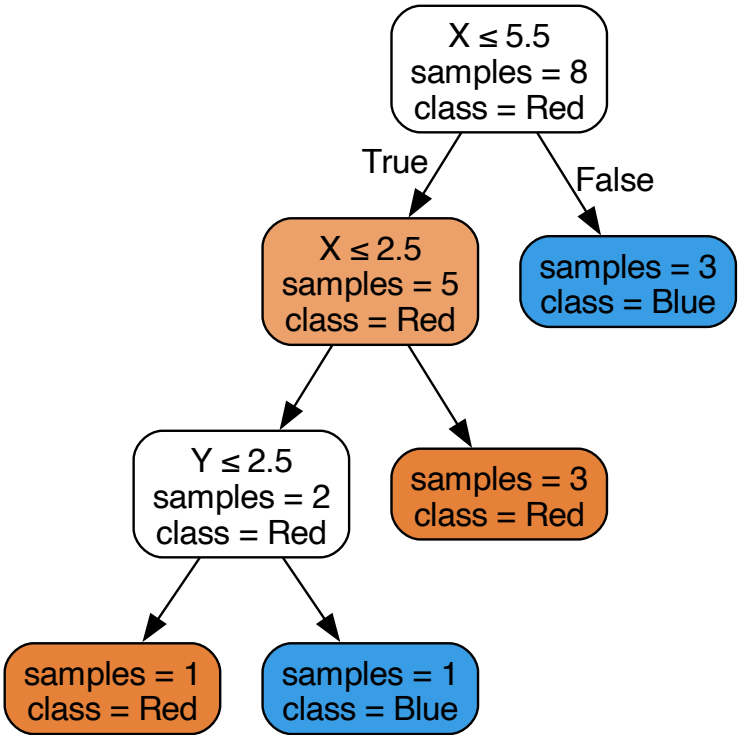


Figure 1.1: An example classification tree matching the partitioning shown above.

tree helped divide the most data, while features referenced only near the tip of the tree only help distinguish a few data points from one another. In this example, we can see that X was a more useful feature for prediction than Y.

In addition to their appealing visual representations as flow-charts, decision trees can produce useful classifiers for a wide range of scenarios. Decision trees do not require a linear relationship between features and predictions, they can utilize both categorical and numeric features, they do not need any normalization of features, and they handle outliers in training data well. This makes them highly appealing to

researchers, who need to do little preparation before deploying a tree.

Random Forests

Decision trees are prone to overfitting: if left unconstrained, they will add more and more decision layers until they subdivide training data into tiny partitions of only one category each. If these partitions are overly specific to the training data, rather than representing patterns that will also be present in testing and real-world data, then tree performance will be poor.

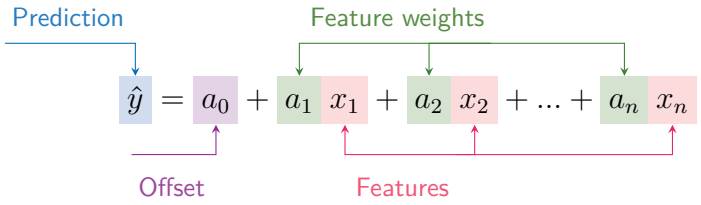
One solution to the overfitting problem is to train several classification trees using a random subset of features. During testing, take a majority vote between trees to predict a new data point's label. Since each tree has access to different features, they will be unable to create the same overfit partitions as one another, and ideally, majority consensus will label data points correctly even if each tree is prone to overfitting. This collection of decision trees is known as a *random forest* [75].

Just as with decision trees, we can analyze the trees in a forest to identify which features most contribute to classifying new data points, giving us a measurement of feature importance.

Random forests are a common classifier choice because they maintain most of the simplicity and interpretability of decision trees, but often yield much better results. For this reason, we utilize a random forest classifier in chapter 2 to distinguish between GitHub and Penumbra git repositories based on a number of features about the repository size and commit contribution history.

Logistic Regression

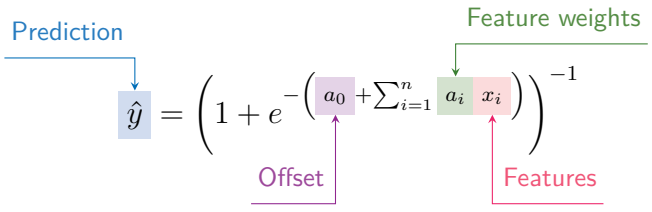
In *linear regression* we fit a line to training data across multiple dimensions using a linear combination of n features, such as:



(1.6)

In machine learning, the coefficients $\vec{a}$ are typically determined numerically based on training data, starting with arbitrary values[1] and then tuning to minimize an objective function like mean-squared error, where the distance between the regression line and each data point constitutes error. The line can now be used for numeric prediction: given features $\vec{x}$ for a test data point, $\hat{y}$ represents a predicted output.

In *logistic regression* we use a similar technique for binary classification rather than numeric prediction. Once again we take a linear combination of features, but we now combine them through a logit function:



(1.7)

The logit function returns a value between 0 and 1 along an S-shaped curve, as

[1] A common choice is setting $\vec{a} = \vec{1}$, or a uniform prior that all features are equally important

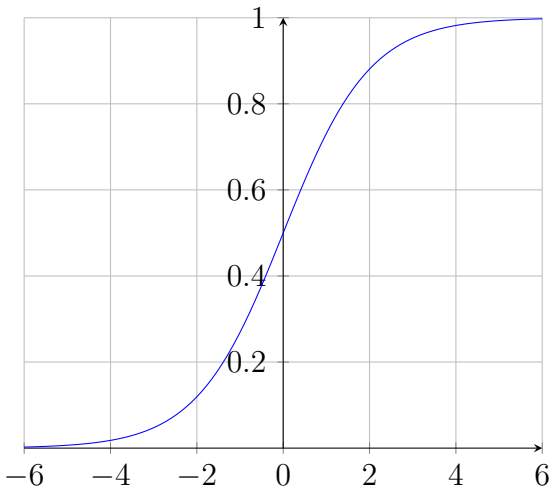


Figure 1.2: Plot of the logit function for only one predictor, or feature. The x-axis represents the input value, and the y-axis represents a "certainty," where 0 represents 100% certainty that the output is in category 1, and 1 represents 100% certainty that the output is in category 2.

shown in fig. 1.2. Here, a $\hat{y}$ of zero represents total certainty that the input data belongs to category 1, a $\hat{y}$ of one represents total certainty that the input belongs to category of 2, and values between zero and one represent a confidence prediction one way or another.

As with linear regression, we choose $\vec{a}$ to minimize error, but here the error is defined in terms of negative log-likelihood for each input k:

$$l_k = \begin{cases} -\ln(\hat{y}) & \text{If } y_k = 1 \\ -\ln(1-\hat{y}) & \text{If } y_k = 0 \end{cases} \tag{1.8}$$

This returns a high error when $\hat{y}$ was confident that an input belonged in the incorrect category, a lower error when $\hat{y}$ is uncertain, and lowest error when $\hat{y}$ is confidently correct.

Logistic regression assumes there is a linear relationship between input features

and output category, and handles outliers poorly. However, it is less prone to overfitting than decision trees, especially in small sample-size scenarios, and is often preferred for simplicity when a clear linear relationship is present.

We utilize logistic regression in chapter 5 when predicting which subreddit a user belongs to based on the language of their comments. In this scenario each feature is the frequency with which a user wrote a particular word, normalized by the prominence of that word across the dataset.

1.2.2 Scoring Classifiers

ROC Curves

Most binary classifiers do not return a single categorical label, but return a confidence, as seen in logistic regression. Even trees can return a confidence score, if the partition selected contains training data from multiple categories. In these scenarios, we must choose a cutoff threshold: do we count a data point as being in category 1 if the classifier is 90% confidence? Is 70% confidence sufficient?

As we tune the cutoff threshold towards 100% we should expect minimal false positives, but many false negatives, as any data points the classifier is not entirely certain about will be misclassified. Likewise, if we tune the threshold towards 0% we will include all correct data in category 1, but we will also include many false positives under the label.

Therefore, the cutoff threshold can be thought of as a compromise between the true positive rate and the false positive rate, where we'd like to find a value that maximizes the former while minimizing the latter. We can visualize this compromise

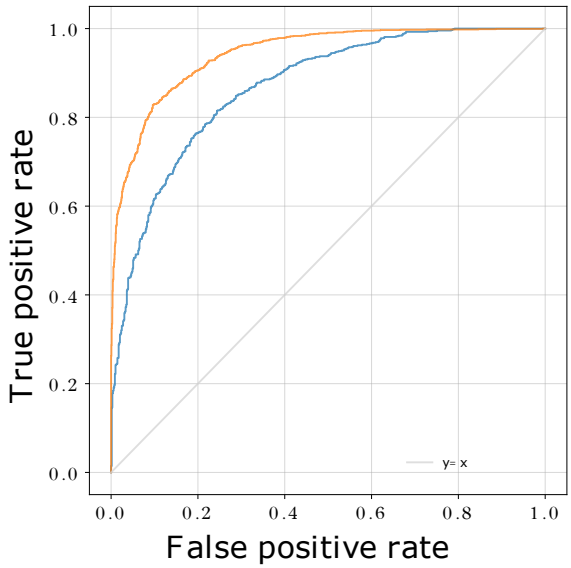


Figure 1.3: An example ROC curve. The orange and blue lines represent two machine-learning classifiers. The orange curve has better performance, because it can achieve a higher true positive rate at a lower false positive rate. This figure is a subset of fig. 5.2, which provides more context.

on a two-dimensional plot called a *receiver operating characteristic* or *ROC* curve. One such curve is shown in fig. 1.3.

As a univariate summary statistic, researchers often report the "area under the curve" (AUC) of a classifier's ROC curve. Here, a value of 1 is optimal, a value of 0.5 is no better than a random guess weighted by category size, and values below 0.5 are worse than guessing. The AUC ignores much of the nuance of an ROC curve and the tune-able behavior of each classifier, but a single number is tempting for directly comparing and ranking classifiers' performance.

Matthews Correlation Coefficient

The Matthews Correlation Coefficient [108], or MCC, is a single-valued summary of classifier performance. It is equivalent to the Phi coefficient [37] and the Yule

coefficient [172]. The MCC is a common measurement of machine-learning classifier quality which accounts for true and false positives and true and false negatives (TP, FP, TN, and FN, respectively). It can be calculated as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The metric is scaled such that $+1$ indicates perfect classification, 0 for results no better than random, and -1 for entirely incorrect classification.

There is contention in the machine-learning community that MCC should supplant ROC AUC as the preferred univalued summary statistic because the latter metric only accounts for accurate measurement of "positive" labels, ignoring the true and false negative rates [30]. These researchers demonstrate that ROC AUC can provide an overly "optimistic" representation of classifier performance. Nevertheless, ROC AUC remains an extremely common metric in studies using classifiers, including my own.

1.3 Social Networks as Graphs

Most of my research focuses on aggregate user behavior, and so my methodology usually centers on analyzing statistical distributions. These distributions are well-suited to measuring simple patterns, such as how active most users are within a community, or how often they use in-group vocabulary, or what a typical team size is in an open-source project. However, networks are ideal for understanding some indirect patterns, such as "is this community a bridge, connecting many users from other diverse communities?" or, "if we removed a particular community from a platform, how would

the remaining population be impacted?" While I do not leverage network analysis in my earlier chapters, they feature prominently in chapter 4. This section provides readers with sufficient background to understand how to represent social networks using graphs, and measure and interpret graph centralization in a variety of ways.

1.3.1 Network Definitions

A network consists of *nodes* (also called *vertices*) that represent people, institutions, or another singular unit of study. *Edges* between nodes (sometimes referred to as *links*) represent a relationship between two nodes. Edges can be *undirected,* representing bidirectional relationships like classmates, or they can be *directed*, indicating a one-directional relationship such as one social media user following another. Additionally, edges can be *unweighted,* indicating only that a relationship exists, or they can be *weighted,* meaning that each edge has a numeric value associated with it indicating the strength of the relationship. For example, a network of financial transactions may indicate both who has paid money to whom, and the amount of money exchanged. Both unweighted, undirected and weighted, directed graphs are illustrated in fig. 1.4.

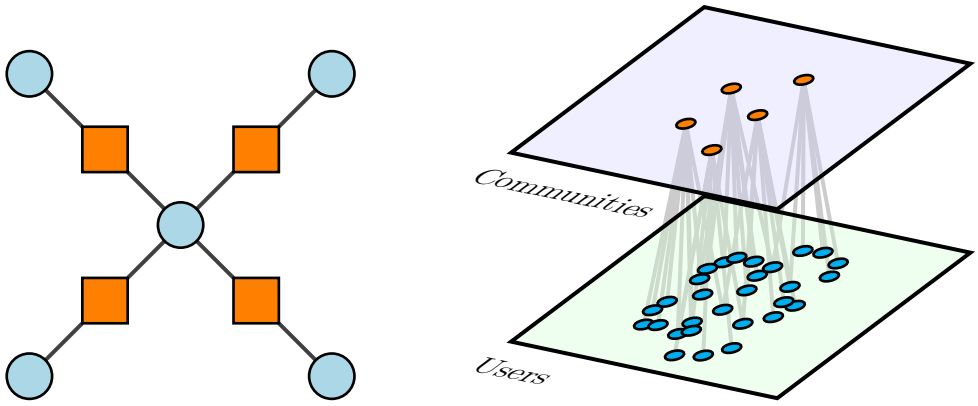
(a) Unweighted, undirected **(b)** Weighted, directed

Figure 1.4: Example networks, with unweighted and undirected edges (left), and weighted and directed edges (right).

Of particular interest in my work are *bipartite* networks. In these two-partition

networks, nodes can be described as belonging to one of two categories, and edges can only exist between nodes in different categories. For example, in a network representing bees and the flowers they help to pollinate, edges may only exist between vertices representing bees and those representing flowers, but edges among bees or among flowers are undefined and not permitted. In my own work, bipartite networks often represent users and communities they interact with. In chapter 4 I use bipartite graphs to represent social users and the communities they participate in, such as Usenet users and the newsgroups they write in.

Bipartite networks can be considered a sub-class of *multilayer* networks, which permit an arbitrary number of node categories, and edges both within and between categories. Two visualizations of bipartite networks are offered in fig. 1.5.



(a) Shapes and colors indicate category **(b)** Layer indicates category

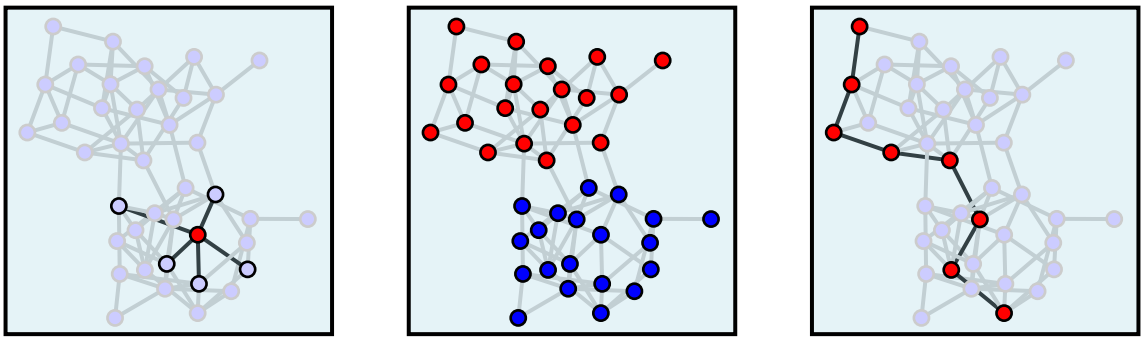
Figure 1.5: Example bipartite networks, with categories distinguished by shape and color (left), and by the positioning on two layers (right).

1.3.2 Centralization

Representing data as a network emphasizes the importance of relationships between elements. We may be interested in the *degree* of a node, or the number of edges it is connected to, and how its degree compares to the distribution of degrees for all vertices in the network. We may be interested in the structural role a node plays in a network, such as whether it acts as a bridge between two regions of a graph, or whether many shortest paths between vertices pass through this particular vertex.

A common goal is to measure how "centralized" a network is. This term can define many different attributes of interest, which broadly fall into three categories illustrated in fig. 1.6: node features, regional features, and global features. Node features describe how well-connected a node is to its peers, such as its degree, or the average path length from a node to other nodes in the graph. These features are often normalized across all nodes, so that we can identify the nodes with the highest degree, or that have the shortest average path lengths. Regional features describe a group of nodes, such as how densely connected nodes are within a group, or their modularity (the ratio of edges within the group to edges leaving the group). These features are often used to describe the behavior of a community, or are used to justify the post-hoc identification of a community. Finally, global features describe the overall topology of a graph. These can include aggregations of node-level features (such as average degree and average shortest-path-length), comparisons to a fully-connected graph (*density* measures the ratio of edges that exist to edges that could exist in a complete graph), or best- and worst-case attributes of the graph (*diameter* measures the longest shortest-path across the graph, or the longest path one might be required

to take to traverse from any node to any other).



(a) Degree **(b)** Clustering **(c)** Diameter

Figure 1.6: Centralization can be defined relative to a single vertex (such as the node's degree, or its average distance from other nodes), groups of vertices (such as a measurement of how insular two clusters are), or the entire graph (such as its diameter, or density).

1.3.3 Network Generating Functions

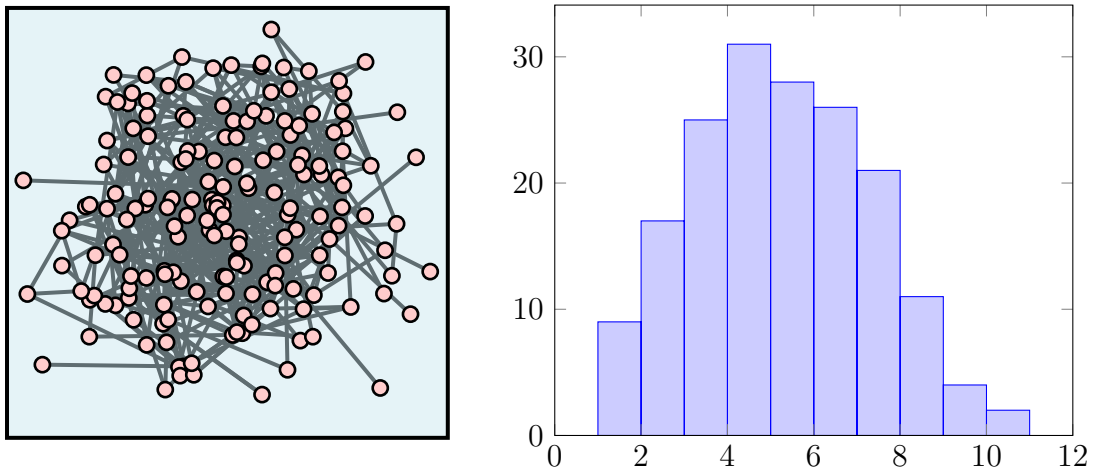
In my research I typically create networks from real-world data, by creating nodes to represent users and communities, and creating edges between users and communities representing some aspect of user interaction data I am interested in studying. However, it is sometimes useful to create *null-model* synthetic networks with a controlled attribute for use as reference. For example, when studying Mastodon, I may create an artificial network with the same number of users and communities, and the same number of edges per community, but with the user side of each edge randomized. By comparing patterns identified on the real-world networks to those found on null-model networks I can determine whether the patterns are explainable as an artifact of the distribution of community sizes, or whether they may be driven by a different aspect of user behavior.

Erdös Rényi

One common type of graph generating function is an Erdös Rényi graph [44], often called an ER-graph or a "random" graph. It is typically modeled formally as $G(n, p)$, implying a graph of n nodes with p probability of an edge between any pair. This can be written in psuedocode as:

```
def G(n,p):
    g = createEmptyGraph(n)
    for i in (0 .. n-2):
        for j in (i+1 .. n-1):
            if( random() < p ):
                createEdge(g, i, j)
    return g
```

ER graphs have, on average, density p and a normal degree distribution with average degree $\langle k \rangle = np$, and an expected $\binom{n}{2}p$ edges. ER graphs are studied frequently because they are very simple to generate, with well-understood properties.



(a) Visualization of an ER network, arbitrary layout showing no clear hubs

(b) Example ER network node degrees, showing a normal distribution

Figure 1.7: Example Erdös-Rényi graph (left) and its approximately normal degree distribution (right)

However, to generate *bipartite* random graphs we must adjust the model to create

two types of nodes, which we will call u for users and c for communities. Edge probability p now describes the probability of creating any possible edge between a user and community. This $G(u, c, p)$ model can be written as:

```
def G(u,c,p):
    g = createEmptyGraph(u+c)
    for user in (0 .. u-1):
        for community in (u .. u+c-1):
            if( random() < p ):
                createEdge(g, user, community)
    return g
```

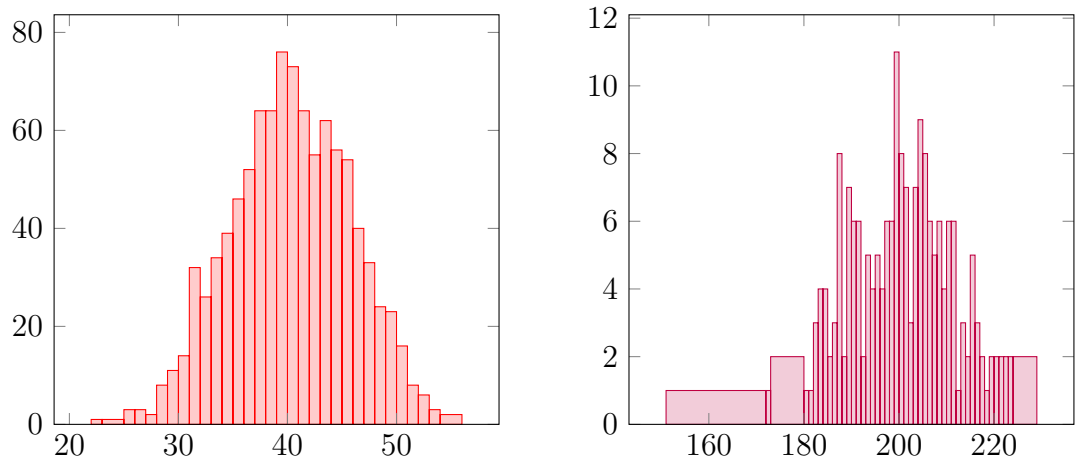
This model has similar properties to typical ER graphs if you examine either node category. Communities have an expected degree $\langle k_c \rangle = up$ while users have an expected degree $\langle k_u \rangle = cp$. The degree distributions of both the users and communities are normal; but the degree distribution of the graph as a whole may appear bimodal, if there is a large difference between the number of users and communities, and therefore a different expected degree for each. This is illustrated in fig. 1.8.

Power-Law Graphs

While ER graphs have normal degree distributions, most social networks exhibit multi-scalar power-law degree distributions. For example, while most Twitter[2] users have few to no followers, a minority of celebrity accounts have millions, a micro-minority of ultra-celebrities have tens of millions, and the top six users have over one hundred million followers each. When creating null-models for social networks it is therefore useful to reproduce similar degree distributions.

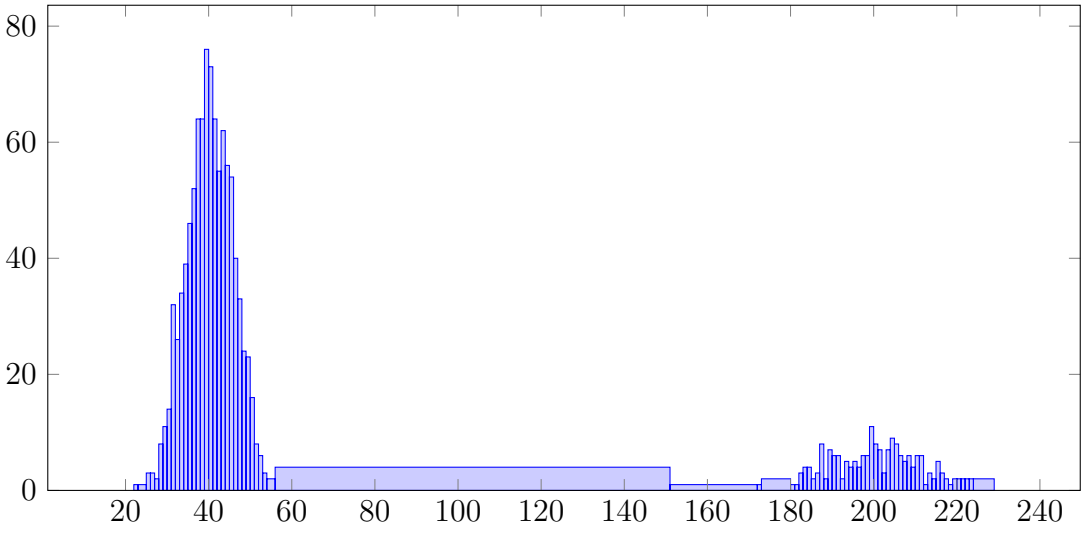
There are two broad strategies for creating graphs with power-law degree distributions. One is to prescribe a desired degree distribution using a Configuration [118] or Chung-Lu Model [33], assigning degrees to nodes by sampling from a desired

[2]Now X



(a) User degree distribution

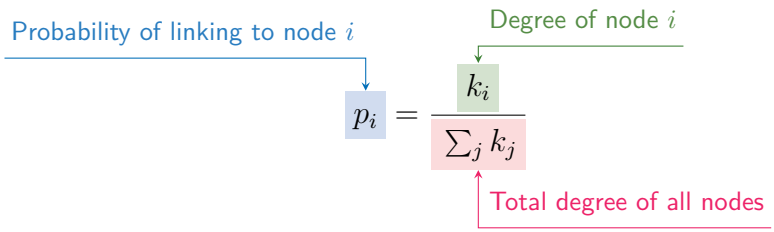
(b) Community degree distribution



(c) Joint degree distribution

Figure 1.8: Degree distributions for a bipartite Erdös-Rényi graph. Both the user and community degrees follow approximate normal distributions, while the overall degree distribution appears bimodal.

power-law curve, then producing a random graph with the assigned degrees. Another strategy is to implement *preferential attachment,* describing some process by which nodes choose to connect to one another. For example, in the Barabási-Albert model [4] (sometimes called the BA-model), each new node connects to m pre-existing nodes, sampled with the following probability:



(1.9)

This generative mechanism means that nodes "prefer" to link to high-degree nodes, creating a *rich-get-richer* pattern wherein early nodes in the network become high-degree hubs. Such preferential attachment mechanisms offer less direct control over degree distributions, but demonstrate how power-law degree distributions can theoretically arrive from simple individual choices.

As with ER graphs, adapting power-law graphs to bipartite settings requires some additional design decisions. Following the prescriptive degree-distribution approach, we can assign a power-law degree distribution to communities, then connect to users uniformly at random until the community degrees are satisfied. This yields a power-law degree distribution for communities, and a normal degree distribution for users. Alternatively, following the BA model, users can select m communities to connect to using preferential attachment. This yields a power-law degree distribution for communities, and a uniform degree distribution for users (that is, all users will have degree m).

1.3.4 Modeling Choices

When modeling real-world interactions as a network, the choice of vertex, edge, and layer definitions is crucial as it both enables and restricts our ability to observe patterns [24]. In the context of social networks, I often use bipartite graphs to represent users and communities they participate in. But what delineates a community, and what constitutes participation? Using Reddit as an example, we may represent users and subreddits as two classes of vertices, where a weighted edge indicates how many comments a user has made on posts within a subreddit. However, this obscures whether the user commented on many posts, perhaps indicating consistent interaction with a community over time, or commented on a single post many times, which does not signify the same group affiliation. We could alternatively define weighted edges as the number of posts in a subreddit that a user has commented on, but this fails to distinguish between a user who has extensive back and forth conversations and one who comments once on many posts. Any network analysis determining which nodes are the most "important," which communities are the most insular, or which graphs are densest, are contingent on these modeling decisions.

Chapter 2

The Penumbra of Open Source

Foreword

Open Source software development is a convenient case study for understanding online group behavior. A software "project" provides a rallying point for a community, and in broad strokes the community members have a shared goal of developing and maintaining the project, even if individuals' objectives within that goal vary widely. The community's activities are predominantly online, recorded, and public. Each project yields an observable artifact, namely the software itself and the version control system (VCS) history documenting code contributions from each member. It is important not to over-emphasize the role of code in software development; many open source contributions are *not* in the form of source code, but in discussion, organization, community management, and other often invisible labor that I have studied elsewhere [171, 111].

Participation in a project is primarily regulated by two factors: the governance structure and policies of the organization, and the technology over which participa-

tion is made. A common framing among open source maintainers and researchers is that open source governance falls into two broad categories [89]: the *benevolent dictator* model, and the *community consensus* model. In the former, a single individual (typically the project founder) makes all final decisions about a project's future, including who can participate in development (or who is made to feel welcome) and what features and design choices will be adopted. This is the default model for most new projects that have not put thought into how they will be governed, but many large projects retain a benevolent dictator, most notoriously the Linux kernel under the stewardship of Linus Torvalds. By contrast, consensus-driven projects have no single leader; members can make small contributions autonomously, and when making larger decisions, they circulate a proposal among the general membership, wait an agreed upon time for discussion or dissenting opinions, then proceed by implied consensus if no objections are raised. There is a wealth of diversity among both camps: there are benevolent dictatorships where the leader intervenes only in the event of stalemate among the membership, or consensus-based projects with a working group and discussion structure bordering on parliamentary.

Another axis along which governance is frequently judged is openness and transparency [133]. Under this framework, *cathedral* projects have an internal development team that periodically releases new versions of the project to the public, but does not share intermediate work or welcome external contributions. By contrast, *bazaar* projects are developed in the open, welcoming new contributions and contributors, typically with a more flexible definition of membership and release versions.

A common lens for understanding governance structure comes from Ostrom's Institutional Analysis and Development framework [121], which in the context of digital

institutions [49] describes three categories of rules: *operational rules,* defining what actions members can take (such as submitting source code or opening a bug report or feature request for discussion), *collective rules*, which describe the shared context in which participants take operational actions and how participants can interact, and *constitutional rules*, the process through which operational, collective, and constitutional rules can be changed.

In the context of open source software development, governance rules are defined by two parties: the leaders of the project, and the developers of the tools they rely on for collaboration. For example, the source code management software *git* has a list of users allowed to contribute to a project. It does not have functionality for requiring democratic approval before contributions are accepted, so any more sophisticated social policies must be awkwardly enforced on top of a less flexible technical layer [174].

In this chapter I examine the impact *GitHub*, the open-source hosting, collaboration, and development platform, has had on open-source software. GitHub has grown to have a central role in the open-source ecosystem: so many projects are hosted on GitHub that it facilitates project discovery for developers, hosts conversations among developers and between developers and users, handles bug reports and project road-mapping, and code review to accept external contributions to a project. GitHub strongly encourages a bazaar model of transparent development and frequent external-contribution. The functionality GitHub offers pre-supposes a particular framing of ownership; for example, GitHub allows "members" of a project to contribute source code directly, while those outside a project must submit a "pull request" to be accepted by a member, while project "owners" have the ability to

designate users as members or co-owners of the project. To better understand the impact of GitHub's framing and functionality on open-source software, I compare the contribution history of projects developed on GitHub, to the *Penumbra*; public projects developed off-platform in GitHub's shadow.

Abstract

GitHub has become the central online platform for much of open source, hosting most open source code repositories. With this popularity, the public digital traces of GitHub are now a valuable means to study teamwork and collaboration. In many ways, however, GitHub is a convenience sample, and may not be representative of open source development off the platform. Here we develop a novel, extensive sample of public open source project repositories outside of centralized platforms. We characterized these projects along a number of dimensions, and compare to a time-matched sample of corresponding GitHub projects. Our sample projects tend to have more collaborators, are maintained for longer periods, and tend to be more focused on academic and scientific problems.

2.1 Introduction

The GitHub hosting platform has long been recognized as a promising window into the complex world of online collaborations [38], open science [126], education [173], public sector work [112], and software development [85]. From 10 million repositories in 2014 [86], GitHub reported over 60 million new repositories in 2020 [58]. How-

ever, despite its size, there remain significant risks associated with GitHub as a data platform [84]. Without a baseline study examining open source development off of GitHub, it is unclear whether public GitHub data is a representative sample of software development practices or collaborative behavior. For studies of collaborations, it is particularly worrisome that most GitHub repositories are private and that most public ones are small and inactive [86]. These data biases have only grown in recent years as the platform stopped limiting the number of private repositories with fewer than four collaborators in 2019 [57].

Despite the fact that GitHub is not a transparent or unbiased window into collaborations, the popularity of the platform alone has proved very attractive for researchers. Early research focused on the value of transparency and working in public, analyzing how individuals select projects and collaborations [38], and conversely how collaborations grow and thrive [88, 115]. While fine-grained information about git commits within code repositories is readily available, higher-level findings about team collaboration and social software development practices are scarcer. Klug and Bagrow [88] introduce a metric of "effective team size," measuring each contributor's contributions against the distribution of GitHub events for the repository, distinguishing peripheral and "drive-by" contributors from more active team members. Choudhary et al. [32] focus on identifying "periods of activity" within a repository, beginning with a simple measurement of time dispersion between GitHub events, then identifying the participants and files edited in each burst of activity to determine productivity and partitioning of work according to apparent team dynamics.

Beyond looking at patterns of collaborations within projects, it is also useful to study GitHub as a social network, where collaborations are social ties mediated by

repository [101, 156, 26]. These studies tend to offer results showing analogies between GitHub collaborations and more classic online social networks, such as modular structure [181] and heterogeneous distributions of collaborators per individual driven by rich-get-richer effects [101, 26]. More interestingly, studies also found that GitHub tends to show extremely low levels of reciprocity in actual social connections [101] and high levels of hierarchical, often star-like groups [181]. There are unfortunately few studies providing context for GitHub-specific findings, and no clear baseline to which they should be compared. Is GitHub more or less collaborative than other platforms of open source development? How much are collaborations shaped by the open source nature of the work, by the underlying technology, and by the platform itself? Altogether, it remains an open problem to quantify just how collaborative and social GitHub is.

GitHub is far from the only platform to host open source projects that use the Git version control system, but it is the most popular. What remains unclear is how much of the open source ecosystem now exists in GitHub's shadow, and how different these open source projects are when compared to their counterpart on the most popular public platforms. To this end, here we aim to study what we call the *Penumbra* of open source: Public repositories on public hosts other than the large centralized platforms (e.g. GitHub, GitLab, Sourceforge and other forges). Specifically, we want to compare the size, the nature and the temporal patterns of collaborations that occur in the Penumbra with that of a comparable random subset of GitHub.

Open source has long been linked to academic institutions [96], including libraries [124, 29], research centers [116, 125], and the classroom [138]. Version control systems such as git have been interesting tools to assist in classroom learning [63, 34],

including computer science [97, 43] and statistics [16] courses. GitHub has played a role in the classroom and for hosting scientific research [173, 46], yet we expect many institutions to be either unwilling or unable to utilize GitHub or other commercial tools [100, 138, 35]. We therefore wish in this work to distinguish between academic and non-academic Penumbra hosts, in order to measure the extent with which academic institutions appear within the Penumbra ecosystem.

The rest of this chapter is organized as follows. In section 2.2 we describe our materials and methods, how we identify and collect Penumbra projects, how we gather a time-matched sample of GitHub projects, and we describe the subsequent analyses we perform on collected projects and the statistical models we employ. We report our results in section 2.3 including our analysis of our Penumbra sample and our comparison to our GitHub sample. Section 2.4 concludes with a discussion of our results, limitations of our study, and avenues for future work.

2.2 Materials and methods

2.2.1 Data collection

We began by identifying various open source software packages that can serve as self-hosted alternatives to GitHub. These included GitLab Community Edition (CE), Gitea, Gogs, cgit, RhodeCode, and SourceHut. We limited ourselves to platforms with a web-git interface similar to mainstream centralized platforms like GitHub and GitLab, and so chose to exclude command-line only source code management like GitoLite, as well as more general project management software like Jitsi and

Phabricator. For each software package, we identified a snippet of HTML from each package's web interface that uniquely identifies that software. Often this was a version string or header, such as `<meta content="GitLab" property="og:site_name">`.

We then turned to Shodan [107] to find hosts running instances of each software package. Shodan maintains a verbose port scan of the entire IPv4 and some of the IPv6 Internet, including response information from each port, such as the HTML returned by each web server. This port scan is searchable, allowing us to list all web servers open to the public Internet that responded with our unique identifier HTML snippets. Notably, Shodan scans only include the default web page from each host, so if a web server hosts multiple websites and returns different content depending on the host in the HTTP request, then we will miss all but the front page of the default website. Therefore, Shodan results should be considered a strict under-count of public instances of these software packages. However, we have no reason to believe that it is a biased sample, as there are trade-offs to dedicated and shared web hosting for organizations of many sizes and purposes.

We narrowed our study to the three software packages with the largest number of public instances: GitLab CE, Gogs, and Gitea. Searching Shodan, we found 59596 unique hosts. We wrote a web crawler for each software package, which would attempt to list every repository on each host, and would report when instances were unreachable (11677), had no public repositories (44863), or required login information to view repositories (2101). We then attempted to clone all public repositories, again logging when a repository failed to clone, sent us a redirect when we tried to clone, or required login information to clone. For each successfully cloned repository, we checked the first commit hash against the GitHub API, and set aside repositories that

matched GitHub content (see section 2.2.4). We discarded all empty (zero-commit) repositories. This left us with 45349 repositories from 1558 distinct hosts.

Next, we wanted to collect a sample of GitHub repositories to compare development practices. We wanted a sample of a similar number of repositories from a similar date range, to account for trends in software development and other variation over time. We chose not to control for other repository attributes, like predominant programming language, size of codebase or contributorship, or repository purpose. We believe these attributes may be considered factors when developers choose where to host their code, so controlling for them would inappropriately constrain our analysis. To gather this comparison sample, we drew from GitHub Archive [62] via their BigQuery interface to find an equal number of "repository creation" events from each month a Penumbra repository was created in. We attempted to clone each repository, but found that some repositories had since been deleted, renamed, or made private. To compensate, we oversampled from GitHub Archive for each month by a factor of 1.5. After data collection and filtering we were left with a time-matched sample of 57914 GitHub repositories.

Lastly, to help identify academic hosts, we used a publicly available list of university domains[1]. This is a community-curated list, and so may contain geographic bias, but was the most complete set of university domains we located.

[1] https://github.com/hipo/university-domains-list

2.2.2 Host analysis

We used geoip lookups[2] to estimate the geographic distribution of hosts found in our Penumbra scan. We also created a simple labelling process to ascertain how many hosts were universities or research labs: Extract all unique emails from commits in each repository, and label each email as academic if the hostname in the email appears in our university domain list. If over 50% of unique email addresses on a host are academic, then the host is labeled as academic. This cutoff was established experimentally after viewing the distribution of academic email percentages per host, shown in the inset of fig. 2.1(c). Under this cutoff, 15% of Penumbra hosts (130) were tagged as academic.

2.2.3 Repository analysis

We are interested in diverging software development practices between GitHub and the Penumbra, and so we measured a variety of attributes for each repository. To analyze the large number of commits in our dataset, we modified git2net [59] and PyDriller [147] to extract only commit metadata, ignoring the contents of binary "diff" blobs for performance. We measured the number of git branches per repository (later, in fig. 2.2, we count only remote branches, and ignore `origin/HEAD`, which is an alias to the default branch), but otherwise concerned ourselves only with content in the main branch, so as to disambiguate measurements like "number of commits."

From the full commit history of the main branch we gather the total number of commits, the hash and time of each commit, the length in characters of each

[2] https://dev.maxmind.com/geoip/geolite2-free-geolocation-data/

commit message, and the number of repository contributors denoted by unique author email addresses. (Email addresses are not an ideal proxy for contributors; a single contributor may use multiple email addresses, for example if they have two computers that are configured differently. Unfortunately, git commit data does not disambiguate usernames. Past work [163, 60] has attempted to disambiguate authors based on a combination of their commit names and commit email addresses, but we considered this out of scope for our work. By not applying identity disambiguation to either the Penumbra or GitHub repositories, the use of emails-as-proxy is consistent across both samples. If identity disambiguation would add bias, for example if disambiguation is more successful on formulaic university email addresses found on academic Penumbra hosts than it is on GitHub data, then using emails as identifiers will provide a more consistent view.) From the current state (head commit of the main branch) of the repository we measure the number of files per repository. This avoids ambiguity where files may have been renamed, split, or deleted in the commit history. We apply *cloc*[3], the "Count Lines of Code" utility, to identify the top programming language per repository by file counts and by lines of code.

We also calculate several derived statistics. The average *interevent time*, the average number of seconds between commits per repository, serves as a crude indicator of how regularly contributions are made. We refine this metric as *burstiness*, a measure of the index of dispersion (or Fano Factor) of commit times in a repository [32]. The index of dispersion is defined as σ_w^2/μ_w, or the variance over the mean of events over some time window w. Previous work defines "events" broadly to encompass all GitHub activity, such as commits, issues, and pull requests. To consistently compare

[3] https://github.com/AlDanial/cloc

between platforms, we define "events" more narrowly as "commits per day". Note that while interevent time is only defined for repositories with at least two commits, burstiness is defined as 0 for single-commit repositories.

We infer the age of each repository as the amount of time between the first and most recent commit. One could compare the start or end dates of repositories using the first and last commit as well, but because we sampled GitHub by finding repositories with the same starting months as our Penumbra repositories, these measurements are less meaningful within the context of our study.

Following Klug and Bagrow [88], we compute three measures for how work is distributed across members of a team. The first, *lead workload*, is the fraction of commits performed by the "lead" or heaviest contributor to the repository. Next, a repository is *dominated* if the lead makes more commits than all other contributors combined (over 50% of commits). Note that all single-contributor repositories are implicitly dominated by that single user, and all two-contributor repositories are dominated unless both contributors have an exactly equal number of commits, so dominance is most meaningful with three or more contributors. Lastly, we calculate an *effective team size*, estimating what the effective number of team members would be if all members contributed equally. Effective team size m for a repository with M contributors is defined as $m = 2^h$, where $h = -\sum_{i=1}^{M} f_i \log_2 f_i$, and $f_i = w_i/W$ is the fraction of work conducted by contributor i. For example, a team with $M = 2$ members who contribute equally ($f_1 = f_2$) would also have an effective team size of $m = 2$, whereas a duo where one team member contributes 10 times more than the other would have an "effective" team size of $m = 1.356$. Effective team size is functionally equivalent to the Shannon entropy h, a popular index of diversity, but is

exponentiated so values are reported in numbers of team members as opposed to the units of h, which are typically bits or nats. Since we only consider commits as work (lacking access to more holistic data on bug tracking, project management, and other non-code contributions [25]), f_i is equal to the fraction of commits in a repository made by a particular contributor. Interpreting the contents of commits to determine the magnitude of each contribution (as in expertise-detection studies like [144]) would add nuance, but would require building parsers for each programming language in our dataset, and requires assigning a subjective value for different kinds of contributions, and so is out of scope for our study. Therefore, the effective team size metric improves on a naive count of contributors, which would consider each contributor as equal even when their numbers of contributions differ greatly.

2.2.4 Duplication and divergence of repositories

It is possible for a repository to be an exact copy or "mirror" of another repository and this mirroring may happen across datasets: a Penumbra repository could be mirrored on GitHub, for example. Quantifying the extent of mirroring is important for determining whether the Penumbra is a novel collection of open source code or if it mostly already captured within, for instance, GitHub. Likewise, a repository may have been a mirror at one point in the past but subsequent edits have caused one mirror to diverge from the other.

Searching for git commit hashes provides a reliable way to detect duplicate repos-

itories, as hashes are derived from the cumulative repository contents[4] and, barring intentional attack [149] on older versions, hash collisions are rare. To determine the novelty of Penumbra repositories, we searched for their commit hashes on GitHub, on Software Heritage (SH), a large-scale archive of open source code [2] and within the Penumbra sample itself to determine the extent of mirroring within the Penumbra. Search APIs were used for GitHub and SH, while the Penumbra sample was searched locally. For each Penumbra repository, we searched for the first hash and, if the repository had more than one commit, the latest hash. If both hashes are found at least once on GitHub or SH, then we have a complete copy (at the time of data collection). If the first hash is found but not the second, then we know a mirror exists but has since diverged. If nothing is found, it is reasonable to conclude the Penumbra project is novel (i.e., independent of GitHub and SH).

To ensure a clean margin when comparing the Penumbra and GitHub samples, we excluded from our analysis (section 2.2.3) any Penumbra repositories that were duplicated on GitHub, even if those duplicates diverged.

2.2.5 Statistical models

To understand better what features most delineate Penumbra and GitHub projects, we employ two statistical models: logistic regression and a random forest ensemble classifier. While both can in principle be used to predict whether a project belongs to the Penumbra or not, our goal here is inference: we wish to understand what features are most distinct between the two groups.

[4]Commit hashes include the files changed by the commit, and the hash of the parent commit, referencing a list of changes all the way to the start of the repository.

For logistic regression, we fitted two models. Exogenous variables were numbers of files, contributors, commits, and branches; average commit message length; average editors per file; average interevent time, in hours; lead workload, the proportion of commits made by the heaviest contributor; effective team size; burstiness, as measured by the index of dispersion; and, for model 1 only, the top programming language as a categorical variable. Given differences in programming language choice in academic and industry [130], we wish to investigate any differences when comparing Penumbra and GitHub projects (see also sections 2.2.1 and 2.3.3). There is a long tail of uncommon languages that prevents convergence when fitting model 1, so we processed the categorical variable by combining Bourne and Bourne Again languages and grouping languages that appeared in fewer than 1000 unique repositories into an "other" category before dummy coding. JavaScript, the most common language, was chosen as the baseline category. Missing values were present, due primarily to a missing top language categorization and/or an undefined average interevent time. Empty or mostly empty repositories, as well as repositories with a single commit, will cause these issues, so we performed listwise deletion on the original data, removing repositories from our analysis when any fields were missing. After processing, we were left with 67,893 repositories (47.26% Penumbra). Logistic models were fitted using Newton-Raphson and odds e^{β} and 95% CI on odds were reported.

For the random forest model, feature importances were used to infer which features were most used by the model to distinguish between the two groups. We used the same data as logistic regression model 2, randomly divided into 90% training, 10% validation subsets. We fit an ensemble of 1000 trees to the training data using default hyperparameters; random forests were fit using scikit-learn v0.24.2. Model perfor-

mance was assessed using an ROC curve on the validation set (see section 1.2.2 for further details). Feature importances were measured with permutation importance, a computationally-expensive measure of importance but one that is not biased in favor of features with many unique values [150]. Permutation importance was computed by measuring the fitted model's accuracy on the validation set; then, the values of a given feature were permuted uniformly at random between validation observations and validation accuracy was recomputed. The more accuracy drops, the more important that feature was. Permutations were repeated 100 times per feature and the average drop in accuracy was reported. Note that permutation importance may be negative for marginally important features and that importance is only useful as a relative quantity for ranking features within a single trained (ensemble) model.

2.3 Results

We sampled the Penumbra of the open-source ecosystem: Public repositories on public hosts independent from large centralized platforms. Our objective is to compare the Penumbra to GitHub, the largest centralized platform, to better understand the representativeness of GitHub as a sample of the open-source ecosystem and how the choice of platforms might influence online collaborations. In section 2.3.1 we begin with an overview of the Penumbra's geographic distribution and the scale of hosts. In section 2.3.2 we analyze the collaboration patterns and temporal features of Penumbra and GitHub repositories. Section 2.3.3 examines the programming language domains of Penumbra and GitHub projects while section 2.3.4 further investigates differences between academic and non-academic Penumbra repositories. Statistical models in

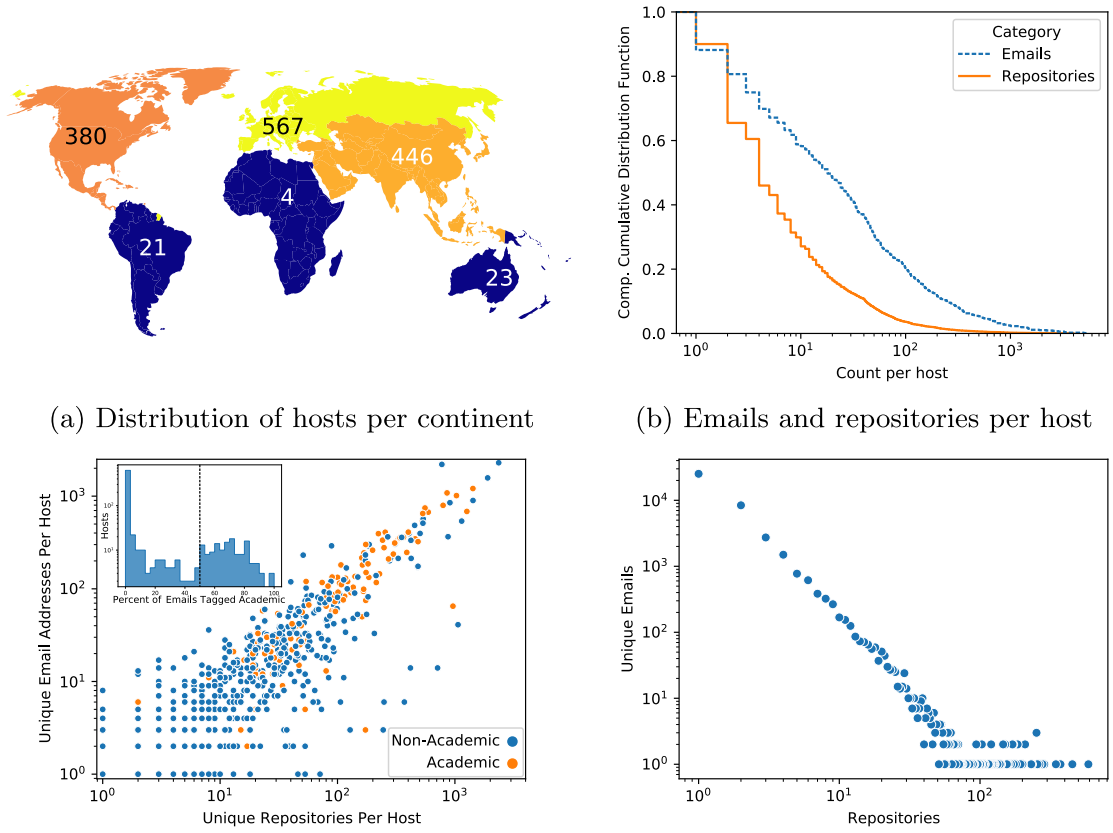
Region	% Hosts	% PN users	% GH users [58]	PN repositories (per capita)	% Unique emails from academic domains
EU	39.35	73.47	26.8	83612 (1.52×10^{-4})	39.20
NA	26.37	15.81	34	51245 (2.97×10^{-4})	41.22
AS	30.95	7.38	30.7	21765 (1.55×10^{-5})	1.21
SA	1.46	1.36	4.9	2776 (1.64×10^{-5})	12.41
OC	1.60	1.93	1.7	3347 (2.58×10^{-4})	65.24
AF	0.28	0.04	2	215 (9.44×10^{-7})	0.00

Table 2.1: Geographic split of our Penumbra (PN) and GitHub (GH) [58] samples.

section 2.3.5 summarize the combined similarities and differences between Penumbra and GitHub repositories. Finally, in section 2.3.6 we investigate the novelty of our Penumbra sample, how many Penumbra repositories are duplicates and whether Penumbra repositories also exist on GitHub and within the Software Heritage [2] archive.

2.3.1 An overview of the Penumbra sample

Our Penumbra sample consists of 1558 distinct hosts from all six inhabited continents and 45349 non-empty repositories with no matching commits on GitHub (section 2.2.4; we explore overlap with GitHub in section 2.3.6). This geographic distribution, illustrated in fig. 2.1 and described numerically in table 2.1, shows that the Penumbra is predominantly active in Europe, North America, and Asia by raw number of hosts and repositories. However, Oceania has the second most repositories per capita, and the highest percentage of academic emails in commits from repositories cloned from those hosts (table 2.1). Overall, the geographic spread of the Penumbra is similar to GitHub's self-reported distribution of users [58], but with a stronger European emphasis and even less Southern Hemisphere representation.



(a) Distribution of hosts per continent

(b) Emails and repositories per host

(c) Comparison by academic use

(d) Repositories per unique email

Figure 2.1: The penumbra of open source. (a) Geographic distribution of hosts and unique email addresses (in parentheses) in our Penumbra sample. (b) Distributions of emails per host and repositories per host. (c) Distribution of unique emails per repositories. (d) Correlation between repositories and emails per host. We see that the number of repositories and email addresses generally correlate, with some outlying hosts with many more repositories than emails. Academic hosts follow the same general trend, except that they tend to be larger than many non-academic hosts. (inset) Hosts are classified as "academic" if over 50 percent of their email addresses end in `.edu` or come from a manually identified academic domain.

We find a strong academic presence in the Penumbra: on 15% of hosts, more than half of email addresses found in commits come from academic domains (see also section 2.3.4). These academic hosts make up many of the larger hosts, but represent a minority of all Penumbra repositories (37% of non-GitHub-mirrors). We plotted the "size" of each host in terms of unique emails and repositories, as well as its academic status, in fig. 2.1(c). We find that while academic hosts tend not to be "small", they do not dominate the Penumbra in terms of user or repository count, refuting the hypothesis that most Penumbra activity is academic.

We are also interested in how distinct hosts are: How many repositories do users work on, and are those repositories all on a single host, or do users contribute to code on multiple hosts? To investigate, we first plot the number of unique email addresses per host in fig. 2.1(b), then count the number of email addresses that appear on multiple hosts. Critically, users may set a different email address on different hosts (or even unique emails per-repository, or per-commit, although this would be tedious and unlikely), so using email addresses as a proxy for "shared users" offers only a lower-bound on collaboration. We find that 91.7% of email addresses in our dataset occur on only one host, leaving 3435 email addresses present on 2-4 hosts. Fifteen addresses appear on 5-74 hosts, but all appear to be illegitimate, such as "you@example.com", emails without domain suffixes like "admin" or "root@localhost", and a few automated systems like "anonymous@overleaf.com". We find 61 email addresses on hosts in two or more countries (after removing fake email addresses by the aforementioned criteria), and 33 on multiple continents (after the same filtering).

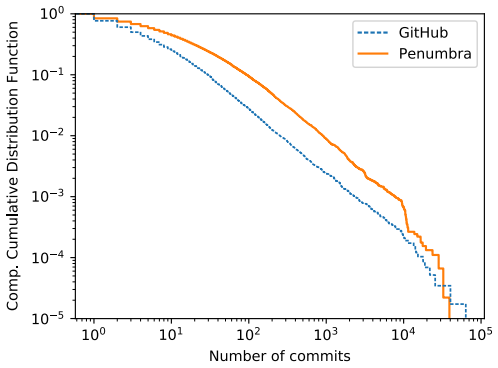
We did not repeat this analysis on our GitHub sample, because the dataset is too different for such a comparison to be meaningful. All GitHub repositories are on a

single “host”, so there is no analogue to “multi-host email addresses”. We considered comparing distributions of “repositories committed to by each email”, but ruled this out because of our data collection methodology. For each Penumbra host, we have data on every commit in every public repository, giving us a complete view of each user's contributions. For GitHub however, we have a small sample of repositories from the entire platform, so we are likely to miss repositories that each GitHub user contributed to.

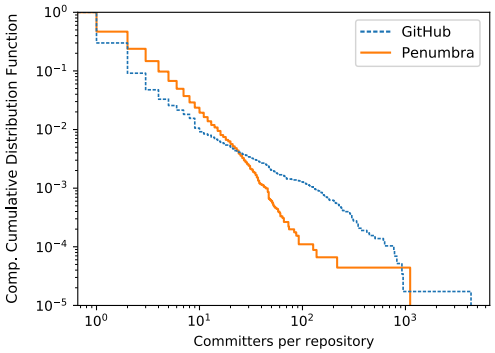
2.3.2 Collaboration patterns and temporal features

We compare software development and collaboration patterns between our Penumbra sample and a GitHub sample of equivalent size and time period (fig. 2.2 and table 2.2). We examine commits per repository, unique emails per repository (as a proxy for unique contributors), files per repository, average editors per file, branches per repository, and commit message length. While mean behavior was similar in both repository samples, diverging tail distributions show that Penumbra repositories usually have more commits, more files, fewer emails, and more editors per file.

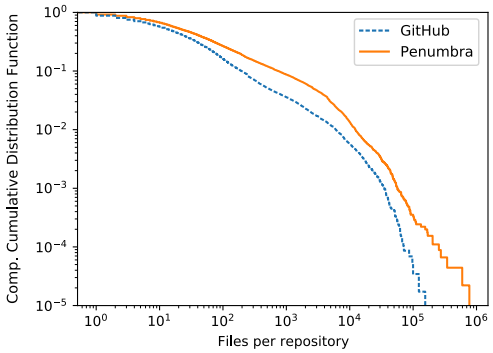
One might hypothesize that with more files and fewer editors the Penumbra would have stronger “partitioning”, with each editor working on a different subset of files. However, our last three metrics suggest that the Penumbra has more collaborative tendencies: while Penumbra repositories are larger (in terms of files), with smaller teams (in terms of editors), there are on average *more* contributors working on the same files or parts of a project. To deepen our understanding of this collaborative



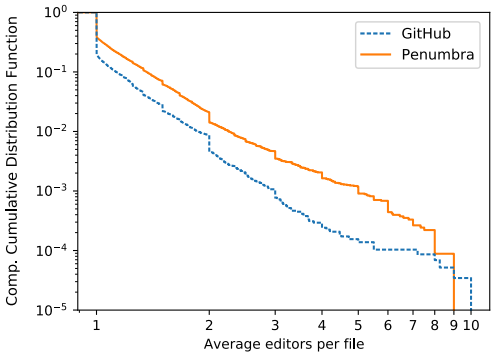
(a) Commits per repository



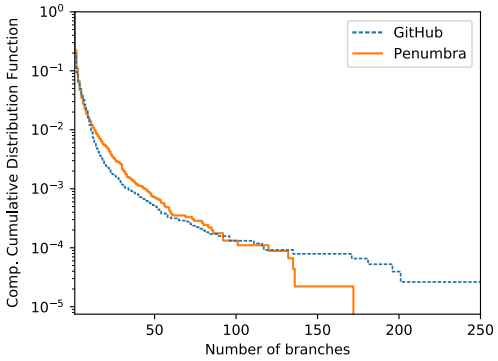
(b) Unique committers per repository



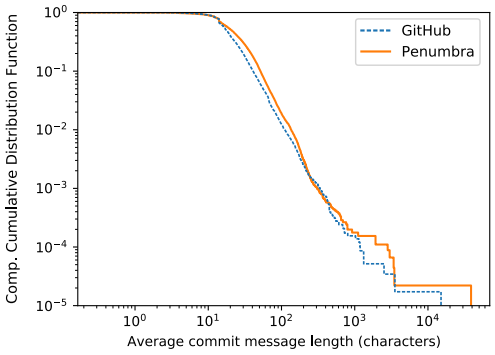
(c) Files per repository



(d) Mean editors per file, per repository



(e) Git branches per repository



(f) Message length per repository

Figure 2.2: Editing and collaborating in the Penumbra and GitHub Comparison of GitHub and Penumbra samples on a variety of metrics. Unique users for all plots are determined by unique email addresses in commit data. File counts are taken at the HEAD commit of the main branch. Editor overlap is defined as average number of unique contributors that have edited each file. Panel (e) excludes two GitHub repositories with 500 and 1300 branches, to make trend comparison easier.

Statistic	Fig.	Penumbra Mean	Median	CI	GitHub Mean	Median	CI	KS 2-Sample KS S	KS P
Files	2.2(c)	244.47	12	[1,859]	156.07	9	[1,264]	0.07	< 0.001
Committers	2.2(b)	2.39	1	[1,6]	2.08	1	[1,3]	0.17	< 0.001
Message Lengths	2.2(f)	29.24	20.80	[7.00,67.33]	24.23	17.60	[7.42,56.00]	0.13	< 0.001
Editor Density	2.2(d)	1.12	1.00	[1.00,1.60]	1.05	1.00	[1.00,1.30]	0.20	< 0.001
Burstiness	2.3(d)	4.86	2.88	[0.50,14.51]	3.68	2.15	[0.17,11.24]	0.13	< 0.001
Commits	2.2(a)	67.12	8	[1,194]	25.27	4	[1,57]	0.20	< 0.001
Branches	2.2(e)	1.74	1	[1,4]	1.67	1	[1,5]	0.03	< 0.001
Age (hours)	2.3(a)	5528	883	[0.1,25556]	2669	73	[0.03,16194]	0.26	< 0.001
Age / Commits	2.3(b)	283	39	[0.02,1261]	193	9	[0.01,944]	0.19	< 0.001
Avg. Interevent	2.3(c)	375	43	[0.05,1547]	257	11	[0.02,1130]	0.19	< 0.001
Team Size	2.3(e)	1.71	1.00	[1.00,3.92]	1.42	1.00	[1.00,2.67]	0.17	< 0.001

Legend: Mean, median, and 5^{th} and 95^{th} percentile values from the Penumbra and GitHub samples for each statistic. **KS S** and **KS P** represent the Kolmogorov-Smirnov two-sample statistic, and its corresponding p-value.

Table 2.2: Comparison of Penumbra and GitHub datasets

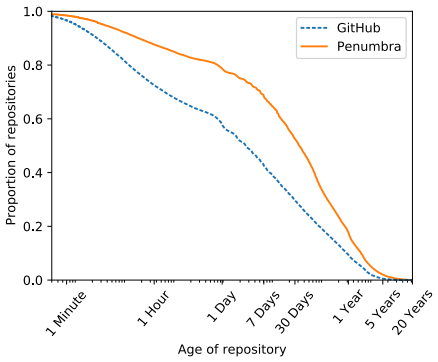
behavior, we also estimated the "effective team size" for each repository by the fraction of commits made by each editor. This distinguishes consistent contributors from editors with only a handful of commits, such as "drive-by committers" that make one pull request, improving upon a naive count of unique emails. These estimates show that while there are more GitHub repositories with one active contributor, and more enormous projects with over 50 team members, the Penumbra has more repositories with between 2 and 50 team members. However, for all team sizes between 2 and 10, we find that more penumbra repositories are "dominated" by a single contributor, see fig. 2.3(f), meaning that their top contributor has made over 50% of all commits.

We also compare temporal aspects of Penumbra and GitHub repositories (fig. 2.3). Penumbra repositories are shown to be generally older in terms of "time between the first and most recent commit" in fig. 2.3(a), have more commits in fig. 2.3(b), but are also shown to have a longer time between commits measured both as interevent time

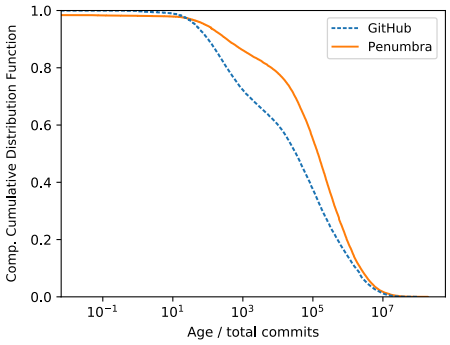
in fig. 2.3(c), and as burstiness in fig. 2.3(d). This means that while Penumbra repositories are maintained for longer (or conversely, there are many short-lived repositories on GitHub that receive no updates), they are maintained inconsistently or in a bursty pattern, receiving updates after long periods of absence. And while both GitHub and Penumbra repositories tend to be bursty, a larger portion of Penumbra repositories exhibit burstiness as indicated by an index of dispersion above 1.

2.3.3 Language domains

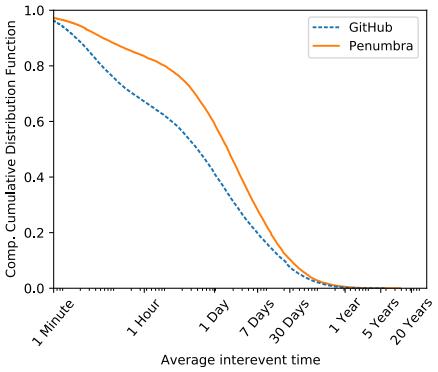
Most of our analysis has focused on repository metadata (commits and files), rather than the content of the repositories. This is because more in-depth content comparison, such as the dependencies used, or functions written within a repository's code, would vary widely between languages and require complex per-language parsing. However, we have classified language prevalence across the Penumbra and GitHub by lines of code and file count per repository in fig. 2.4(left column). We find that the Penumbra emphasizes academic languages (TeX) and older languages (C, C++, PHP, Python), while GitHub represents more web development (JavaScript, TypeScript, Ruby), and mobile app development (Swift, Kotlin, Java). We additionally compare repositories within the Penumbra that come from academic hosts ($> 50\%$ emails come from academic domains; see Methods) and non-academic hosts, using the same lines of code and file count metrics in fig. 2.4(right column). Academic hosts unsurprisingly contain more languages used in research (Python, MATLAB, and Jupyter notebooks), and languages used in teaching (Haskell, assembly, C). Despite Java's prevalence in enterprise and mobile app development, and JavaScript's use in web development, academic hosts also represent more Java and Typescript development. By contrast,



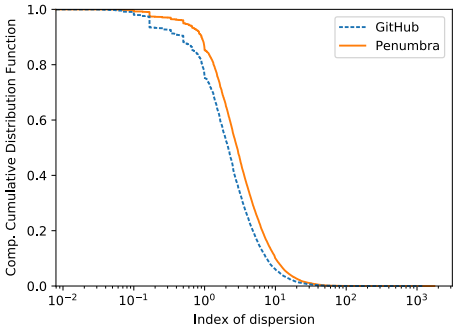
(a) Age per repository



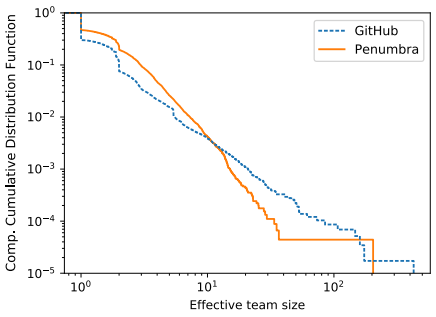
(b) Age / # commits, per repository



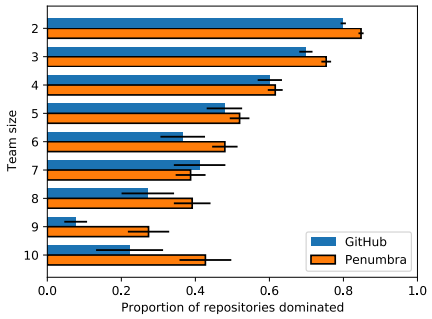
(c) Mean interevent time per repository



(d) Burstiness per repository



(e) Estimated team size by commit distribution, per repository



(f) Percent of repositories dominated by a single committer

Figure 2.3: Temporal characteristics of collaboration in the Penumbra and GitHub. Comparison of GitHub and Penumbra samples on a variety of temporal metrics. The age of repository is given by the time between the first and latest commit. Panels (b-d) look at the heterogeneity of time between events. We first compare the distribution of mean interevent time to the distribution of ratios of age to number of commits, then show the distribution of index of dispersion per repository. Panels (e-f) compare how collaborative repositories actually are, or whether they are dominated by a single committer.

non-academic hosts contain more desktop or mobile app development (Objective C, C#, QT), web development (JavaScript, PHP), shell scripts and docker files, and, surprisingly, Pascal.

2.3.4 Academic and non-academic hosts

Academic hosts account for over 15% of hosts and 37% of repositories in the Penumbra, so one might hypothesize that academic software development has a striking effect on the differences between GitHub and the Penumbra. To investigate this, fig. 2.5 redraws figs. 2.2d and 2.3a with academic and non-academic Penumbra repositories distinguished. We find that the academic repositories are maintained for about the same length of time as their non-academic counterparts, and that academic repositories have fewer editors per file than non-academic development. In fact, academic repositories more closely match GitHub repositories in terms of editors per file. Therefore, we find that academic software development does not drive the majority of the differences between GitHub and the Penumbra.

2.3.5 Statistical models

To understand holistically how these different features delineate the two data sources, we perform combined statistical modeling. First, we performed logistic regression (table 2.3) on the outcome variable of GitHub vs. Penumbra, see section 2.2.5 for details. We fit two models, one containing the primary programming language as a feature and the other not. Examining the odds e^{β} for each variable, we can determine which variables, with other variables held constant, most clearly distinguish GitHub and

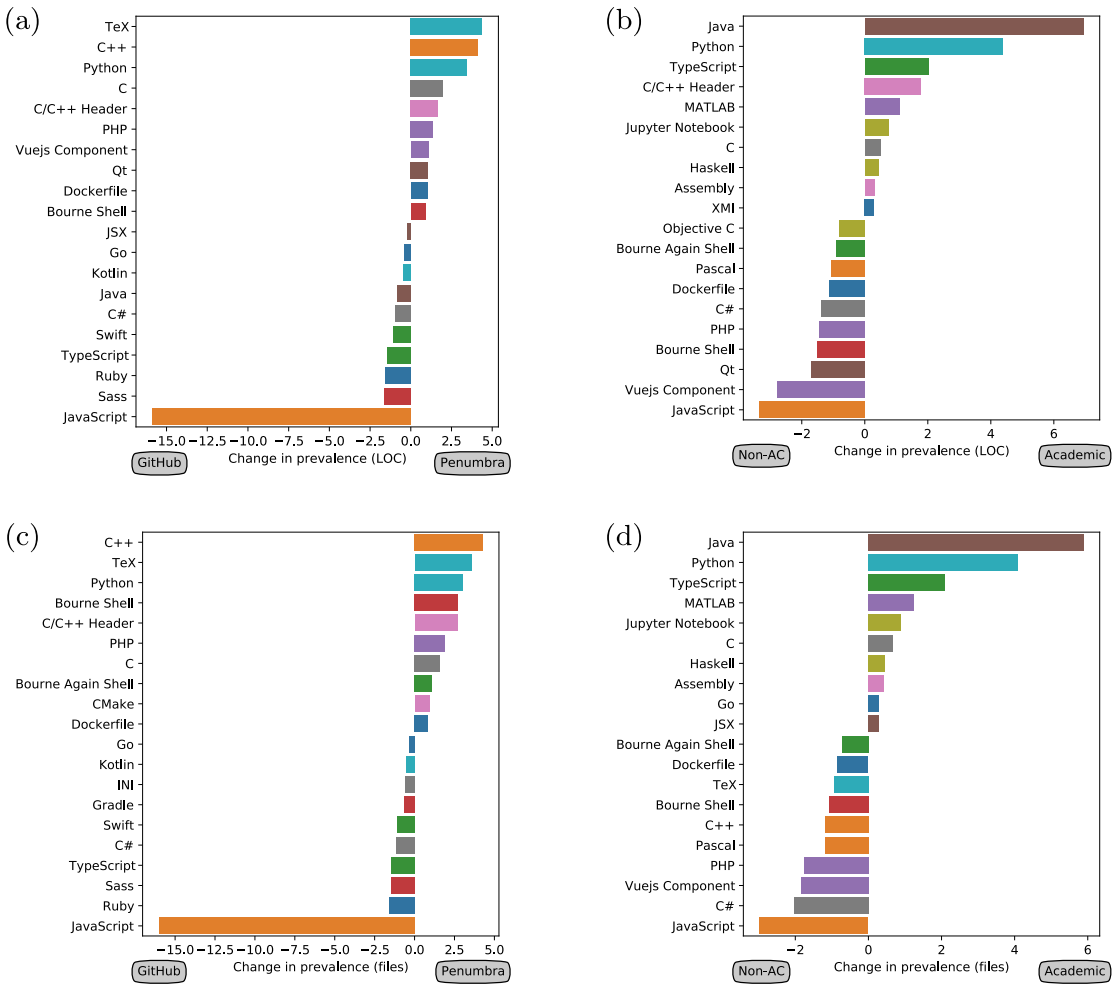


Figure 2.4: Dominant language domains in the Penumbra and GitHub Comparison of language popularity, measured by lines of code (LOC) in panels (a-b) and by file count in panels (c-d). We count the top languages of each repository by the specified metric, normalize the results to a percentage of independent or GitHub repositories, then subtract the percentages. Therefore a language with a value of -0.05, for example, is the top language on 5% more GitHub repositories than Penumbra repositories, while a positive value indicates 5% more Penumbra repositories than GitHub repositories.

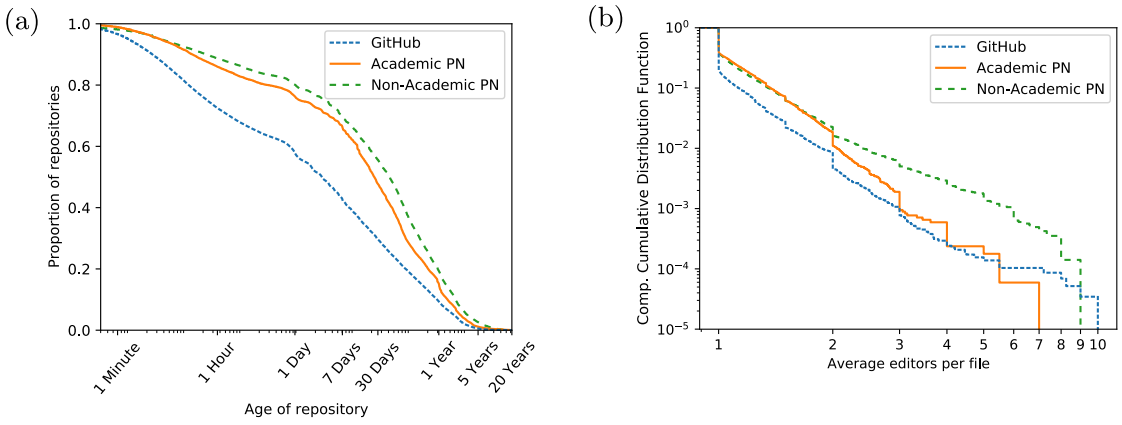


Figure 2.5: Comparing academic and non-academic Penumbra repositories to GitHub Fifteen percent of Penumbra hosts are "academic" under our definition, representing 37% of all Penumbra repositories. We find that academic repositories are maintained for about as long as non-academic Penumbra repositories, so academic development practices do not drive the divergence from GitHub development patterns that we observe. Academic repositories have fewer editors per file than non-academic Penumbra repositories, however, more closely matching development practices seen on GitHub. This refutes the hypothesis that the Penumbra differs widely from GitHub primarily due to academic influence.

Penumbra repositories. The strongest non-language separators are average editors per file, lead workload, and the number of contributors. The strongest language separators are TeX, C/C++ Headers, and C++. The odds on these variables underscore our existing results: Penumbra projects have more editors per file and less workload placed upon the lead contributor. Likewise, the odds on TeX and C/C++ code make it more likely for Penumbra projects to be focused on academic and scientific problems.

Supplementing our logistic models we also used nonlinear random forest regressions trained to predict whether a project was in GitHub or the Penumbra. While trained models can be used as predictive classifiers, our goal is to interpret which model features are used to make those predictions, so we report in fig. 2.6 the top-ten feature importances (section 2.2.5) in our model. Here we find some differences and

	Model 1			Model 2		
	e^{β}	p	CI	e^{β}	p	CI
Constant	0.188	< 0.001	[0.156,0.225]	0.435	< 0.001	[0.364,0.520]
Language (vs. JavaScript)						
Bourne (Again) Shell	3.478	< 0.001	[3.162,3.826]			
C	4.065	< 0.001	[3.671,4.502]			
C#	1.589	< 0.001	[1.444,1.750]			
C++	5.636	< 0.001	[5.184,6.127]			
C/C++ Header	5.829	< 0.001	[5.103,6.657]			
Java	2.192	< 0.001	[2.070,2.321]			
Jupyter Notebook	2.722	< 0.001	[2.459,3.012]			
OTHER	2.124	< 0.001	[2.023,2.230]			
PHP	2.524	< 0.001	[2.323,2.743]			
Python	2.804	< 0.001	[2.651,2.965]			
TeX	30.641	< 0.001	[25.348,37.040]			
TypeScript	1.078	0.187	[0.964,1.205]			
Vuejs Component	4.940	< 0.001	[4.331,5.635]			
Files	1.000	< 0.001	[1.000,1.000]	1.000	< 0.001	[1.000,1.000]
Commits	1.001	< 0.001	[1.001,1.001]	1.001	< 0.001	[1.001,1.001]
Average editors per file	3.337	< 0.001	[3.002,3.709]	3.328	< 0.001	[3.002,3.689]
Average message length	1.002	< 0.001	[1.001,1.002]	1.002	< 0.001	[1.001,1.003]
Burstiness	1.059	< 0.001	[1.055,1.063]	1.058	< 0.001	[1.053,1.062]
Average interevent time [h]	1.000	< 0.001	[1.000,1.000]	1.000	< 0.001	[1.000,1.000]
Branches	0.971	< 0.001	[0.965,0.976]	0.961	< 0.001	[0.955,0.966]
Lead workload	0.461	< 0.001	[0.407,0.522]	0.433	< 0.001	[0.384,0.489]
Committers	0.945	< 0.001	[0.936,0.955]	0.940	< 0.001	[0.931,0.950]
Effective team size	1.010	0.516	[0.980,1.040]	1.016	0.315	[0.985,1.047]
-2LL	84478.844			89788.631		
Pseudo-R2	0.100			0.044		

Table 2.3: Logistic regression models for GitHub vs. Penumbra outcome.

similarities with the (linear) logistic regression results. Both average editors per file and number of contributors were important, but the random forest found that lead workload was not particularly important. However, the most important features for the random forests were average interevent time, burstiness, and number of commits. (All three were also significant in the logistic regression models.) The overall predictive performance of the random forest is reasonable (fig. 2.6 inset). Taken together, the random forest is especially able to separate the two classes of projects based on

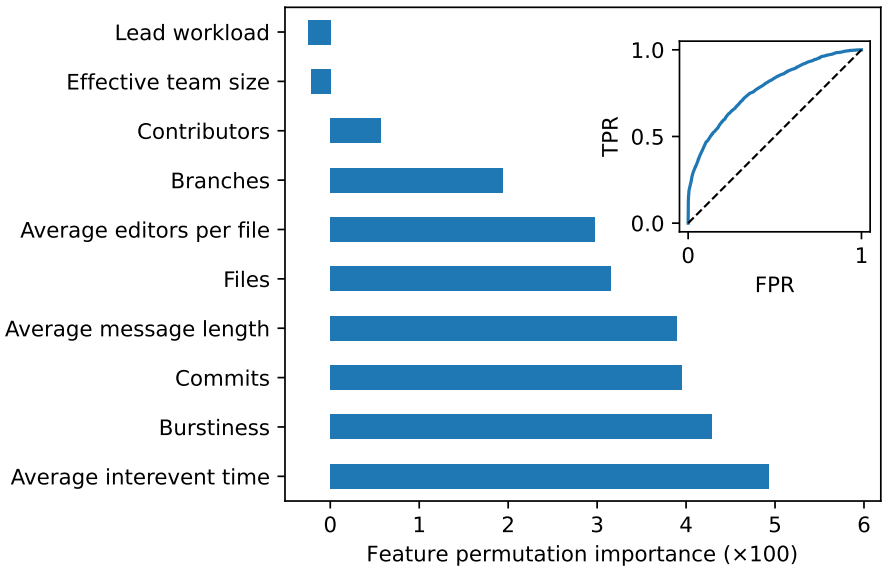


Figure 2.6: Random forest model to delineate Penumbra and GitHub samples. Feature permutation importance (section 2.2.5) once nonlinear random forest regressions were trained to predict whether a project was on GitHub or in the Penumbra. The predictive performance is shown in the inset using an ROC curve of true positive rate (TPR) and false positive rate (FPR).

time dynamics.

2.3.6 Novelty of the Penumbra sample

How novel are the repositories we have discovered in the Penumbra? It may be that many Penumbra repositories are "mirrored" on GitHub, in which case the collected Penumbra sample would not constitute especially novel data. In contrast, if few repositories appear on GitHub, then we can safely conclude that the Penumbra is a novel collection of open source code. To test the extent that the Penumbra is independent of GitHub, we checked the first commit hash of each Penumbra repository against the GitHub Search API (section 2.2.4). We found 9994 such repositories (fig. 2.7) and

conclude that the majority of Penumbra repositories are novel. We excluded these overlapping repositories from our comparisons between the Penumbra and GitHub. However, such repositories may not represent true duplicates, but instead "forks", where developers clone software from GitHub to the Penumbra and then make local changes, or vice-versa, leading to diverging code. To disambiguate, we checked the *last* commit hash from each of the 9994 overlapping repositories against the GitHub API, and found 3056 diverging commits, as illustrated in fig. 2.7. In other words, 30% of Penumbra repositories with a first commit on GitHub also contain code *not* found on GitHub. While we still excluded these repositories to ensure a wide margin between the samples, in fact, the differences in these repositories further underscore the novelty of the Penumbra data.

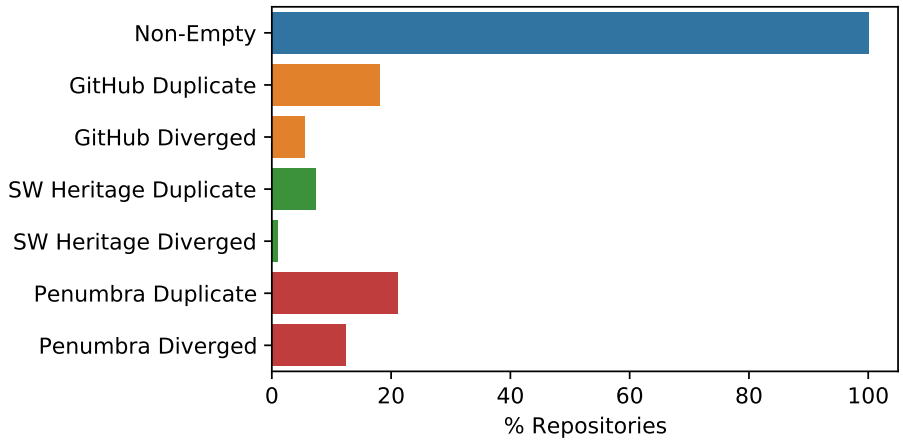


Figure 2.7: The Penumbra's intersection with other datasets. Of the 55343 discovered, non-empty repositories, 18% have a first commit hash that can also be found on GitHub (GitHub Duplicate), but 30% of those repositories diverge and contain code not found on GitHub (GitHub Diverged). Likewise, 7% of Penumbra repositories have a first commit archived by Software Heritage [2], and 14% of those contain code not archived by Software Heritage. Finally, 21% of Penumbra repositories share a commit with one or more other Penumbra repositories, and of these, 58% have unique final commits.

We also compared our repositories against Software Heritage [2], an archive of open

source software development. While Software Heritage is not a hosting platform like GitHub, it represents a potentially similar dataset to our own. Applying the same methodology as for GitHub mirror detection, we found that 4053 repositories (9% of our non-empty Penumbra sample) had a matching first commit hash archived on Software Heritage, and that of these, 564 repositories (14% of overlapping first commits) contained code *not* archived by Software Heritage. Since Software Heritage is an archive, rather than a software development platform, we did not filter out the 4053 overlapping repositories from our comparisons between the Penumbra and GitHub. We again conclude that our Penumbra sample is primarily not captured by Software Heritage; see also Discussion.

We additionally looked for mirrors and forks within the Penumbra, shown in fig. 2.7. As when comparing to external datasets, we found repositories that shared a first commit hash, then checked whether the last commit hash diverged. We find 11717 Penumbra repositories share a first commit with at least one other, which constitutes 25.88% of non-empty Penumbra repositories. These mirrors come from a total of 3348 initial commits. Of these repositories, 6806 share a *last* commit with one or more repositories, suggesting that they have not diverged since creation. Notably, 1287 of the forks and mirrors contain only a single commit. Over a third of the forks and mirrors are on academic hosts (39.46%, 4623 repositories), which is especially notable because academic hosts constitute only 15% of our dataset. As a ratio, we find 35.56 mirrors per academic host, 9.98 per non-academic host. This would fit an educational use-case, such as a class assignment where each student clones an initial repository and then works independently.

Because we are comparing commit hashes, we cannot detect duplicate repositories

if file contents are copied in a history-destructive manner. For example, if someone downloaded the files from a GitHub repository without cloning the git history, then created a *new* repository in the Penumbra with those file contents, the commit hashes will *not* match. This is not how open source projects are typically forked or mirrored, and we have no reason to suspect that this is commonplace in the Penumbra dataset, but it is an important caveat to our methodology.

2.4 Discussion

In this chapter, we collected independent git hosts to sample what we call the *Penumbra* of the open source ecosystems: public hosts outside of the large, popular, centralized platforms like GitHub. Our objective was to compare a sample of the Penumbra to GitHub to evaluate the representativeness of GitHub as a data source and identify the potential impact of a platform on the work it hosts. In doing so, we found that projects outside of centralized platforms were more academic, longer maintained, and more collaborative than those on GitHub. These conclusions were obtained by looking at domains of email addresses of user accounts in the repositories, as well as measuring temporal and structural patterns of collaborations therein.

Importantly, projects in the Penumbra also appear to be more heterogeneous in important ways. Namely, we find more skewed distributions of files per repository and average number of editors per file, as well as more bursty patterns of editing. These bursty patterns are characterized by a skewed distribution of interevent time; meaning, projects in the Penumbra are more likely to feature long periods without edits before periods of rapid editing. Altogether, our results could suggest that the

popularity and very public nature of GitHub might contribute to a large amount of low-involvement contributors (or so-called "drive-by" contributions).

Our current sample of the Penumbra is extensive, but our methodology for identifying hosts presents shortcomings. Most notably, of the approximately 60,000 GitLab CE, Gitea, and Gogs instances identified by Shodan, only 13.4% provided public access to one or more repositories. We can say little about the hosts that provide no public access, and therefore constitute the dark shadow of software development. Further, Shodan may not capture all activity on a given server: it identifies hosts by their responses to a request for the front page, and is not a complete web crawler (section 2.2.1). While this was sufficient in identifying 60,000 hosts, it is an underestimate of the true number of Penumbra hosts, meaning that our dataset remains a sample of the full Penumbra and there exists room for improvement.

We determined from commit hashes that our sample of the Penumbra is mostly disjoint from GitHub and from the Software Heritage archive. This shows that our strategy of seeking public hosts using Shodan is a viable way to uncover novel sources of code. Archival efforts such as Software Heritage and World of Code [105] can benefit from this work as they can easily integrate our sampling method into their archiving process. Doing so can help them further achieve their goals of capturing as much open source software as possible.

There remain several open questions about our sample of the Penumbra worth further pursuit. For instance, the observed shift in languages used on Penumbra repositories implies that they tend to have more focus on academic and/or scientific projects than GitHub. However, programming language alone is a coarse signal of the intent and context of a given project. Future work should attempt a natural language

analysis of repository contents to better identify the type of problems tackled in different regions of the open-source ecosystem. Furthermore, this would allow researchers to match Penumbra and GitHub repositories by the problem-spaces they address, indicating whether developers off of GitHub solve similar problems in different ways.

There are also several important demographic questions regarding, among others, the age, gender, and nationality of users in the Penumbra. GitHub is overwhelmingly popular in North America [58] and therefore does not provide uniform data on members of the open-source community. Critical new efforts could attempt to assess the WEIRDness — i.e., the focus on Western, educated, industrialized, rich and democratic populations [70, 71] — of GitHub as a convenience sample.

Digging further into the code or user demographics of the Penumbra would allow us to answer new questions about the interplay of code development with the platform that supports it. How does the distribution of developer experience levels affect projects, teams and communities? What are the key differences in intent, practices and products based on how open and public the platform is? Who contributes to the work and does it differ depending on the platform [25]?

We are only beginning to explore the space of open source beyond GitHub and other major central platforms. The Penumbra hosts explored here are fundamentally harder to sample and analyze. The hosts themselves have to be found and not all hosts provide public access. Unlike GitHub, we do not have a convenient API for sampling the digital traces of collaborations, so the underlying git repositories must be analyzed directly. There is therefore much of the open source ecosystem left to explore. Yet only by exploring new regions, as we did here, can we fully understand how online collaborative work is affected by the platforms and technologies that support it.

Chapter 3

When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban

Foreword

In chapter 2, I focus on the downstream social effects of technical decisions, specifically the functionality offered by GitHub. In this chapter, I focus on the social impacts of an explicit governance policy: In 2020, Reddit Incorporated chose to revise their stance on acceptable content on the site, and banned approximately 2000 subreddits, or user-governed communities, for harassment and hate speech. This

was a significant reversal from their previous stance of "free-speech absolutism[1]" and a hands-off approach to allowing communities to set their own acceptable content policies enforced by volunteer moderators. Reddit did not ban the users from these subreddits, but only the communities themselves, allowing users to continue participating in other community spaces. This provides an unusual opportunity to observe how users respond to community-removal.

In this study, I led a team of interdisciplinary scholars to study aggregate user behavior after community bans, observing changes in activity levels and in-group vocabulary usage across both typical community-members and the most active "power users." The effects differ between subreddits, and so we examined the kinds of communities that were banned, and how community response to a ban correlated with the category of community.

Importantly, by studying Reddit we can only observe the impact on users who chose to remain active on Reddit. A primary effect of deplatforming is moving content *off* of a platform. We expect, and in some incidences know, that many communities responded to these bans by migrating to other platforms, primarily Voat and the .Win network. While cross-platform behavioral changes are out of scope for this chapter, I make small steps towards such a comparison in chapters 4 and 5, and address this topic further in the conclusion.

[1] https://www.theatlantic.com/technology/archive/2020/12/reddit-ovarit-the-donald/617320/

Abstract

Community-level bans are a common tool against groups that enable online harassment and harmful speech. Unfortunately, the efficacy of community bans has only been partially studied and with mixed results. Here, we provide a flexible unsupervised methodology to identify in-group language and track user activity on Reddit both before and after the ban of a community (subreddit). We use a simple word frequency divergence to identify uncommon words overrepresented in a given community, not as a proxy for harmful speech but as a linguistic signature of the community. We apply our method to 15 banned subreddits, and find that community response is heterogeneous between subreddits and between users of a subreddit. Top users were more likely to become less active overall, while random users often reduced use of in-group language without decreasing activity. Finally, we find some evidence that the effectiveness of bans aligns with the content of a community. Users of dark humor communities were largely unaffected by bans while users of communities organized around white supremacy and fascism were the most affected. Altogether, our results show that bans do not affect all groups or users equally, and pave the way to understanding the effect of bans across communities.

3.1 Introduction

Online spaces often contain toxic behaviors such as abuse or harmful speech [20, 141, 79, 140, 64, 137, 53, 148, 123, 145, 99]. Such toxicity may result in platform-wide decreases in user participation and engagement which, combined with external pressure

(e.g., bad press), may motivate platform managers to moderate harmful behavior [140, 64]. Moreover, the radicalization of individuals through their engagement with toxic online spaces may have real-world consequences, making toxic online communities a cause for broader concern [120, 64, 137, 136].

Reddit is a social media platform that consists of an ecosystem of different online spaces. As of January 2020, Reddit had over 52 million daily active users organized in over 100,000 communities, known as "subreddits", where people gather to discuss common interests or share subject- or format-specific creative content and news [134]. Every post made on Reddit is placed in one distinct subreddit, and every comment on Reddit is associated with an individual post and therefore also associated with a single subreddit. As Reddit continues to gain popularity, moderation of content is becoming increasingly necessary. Content may be moderated in several ways, including: (1) by community voting that results in increased or decreased visibility of specific posts, (2) by subreddit-specific volunteer moderators who may delete posts or ban users that violate the subreddit guidelines, and (3) by platform-wide administrators that may remove posts, users, or entire communities which violate broader site policies. The removal of an entire subreddit is known as a "subreddit ban," and does not typically indicate that the users active in the subreddit have been banned.

Given that the ostensible purpose of subreddit bans is to remove subreddits that are in habitual noncompliance with Reddit's Terms of Service, it is important to understand whether such bans are successful in reducing the offending content. This is especially of interest when the offending content is related to harmful language. Though limited, there is some evidence to suggest that subreddit bans may be effective by certain metrics. Past work has demonstrated that these bans can have both user-

and community-level effects [68, 27, 140, 137, 155, 64]. Several of these studies have suggested that (1) subreddit bans may lead a significant number of users to completely stop using the site, and that (2) following a ban, users that remain on the platform appear to decrease their levels of harmful speech on Reddit [140, 155, 64]. Chandrasekharan et al. [27] also illustrated that postban migrations of users to different subreddits did not result in naive users adopting offensive language related to the banned communities. More work is required to better understand changes in the language of individual users after such bans.

3.2 Previous work

Previous research provides a foundation for investigating the effects of subreddit bans on harmful language and user activity. Detection of offensive content typically takes the form of automated classification. Different machine learning approaches have been applied with varied success, including but not limited to support vector machines and random forests to convolutional and recurrent neural networks [175, 22, 54, 91, 106, 128, 67, 165, 179]. More recently, Garland et al. [52] used an ensemble learning algorithm to classify both hate speech and counter speech in a curated collection of German messages on Twitter. Unfortunately, these approaches require labeled sets of speech to train classifiers and therefore risk not transferring from one type of harmful speech (e.g. misogyny) to another (e.g. racism). We therefore aim for a more flexible approach that does not attempt to classify speech directly, but rather identifies language over-represented in harmful groups; i.e., their in-group language. That language is not a signal of, for example, hate speech per se. In fact, any group is

likely to have significant in-group language (e.g. hockey communities are more likely to use the word "slapshot"). However, detection of in-group language can be fully automated in an unsupervised fashion and is tractable.

The majority of past work on bans of harmful communities on Reddit only examined one or two subreddits, often chosen due to notoriety [68, 27, 140, 137, 64, 155]. Many of these studies focused on the average change in behavior across users and did not consider the factors which may drive inter-individual differences in behavior following a ban [27, 140, 64]. Different users may respond differently to subreddit bans based on their level of overall activity or community engagement. For example, Ribeiro et al. [137] found that users that were more active on Reddit prior to a subreddit ban were more likely to migrate to a different platform following a ban. A user's activity levels prior to a ban also impacted whether activity levels increased or decreased upon migrating to a different platform [137]. Similarly, Thomas et al. [155] demonstrated that users who were more active in a subreddit prior to a ban were more likely to change their behavior following the banning of that subreddit, but the authors did not investigate the ways in which users changed their behavior. Lastly, Hazel Kwon and Shao [68] found that a user's pre-ban activity level within r/alphabaymarket influenced post-ban shifts in communicative activity.

While we are interested in the effects of moderation on any online community, we study Reddit because the platform is strongly partitioned into sub-communities, and historical data on both subreddits and users are readily available [15]. Reddit users are regularly active in multiple subreddits concurrently, and unlike other sub-community partitioned platforms like Discord, Slack, or Telegram, we can easily retrieve a user's activity on *all* sub-communities. This provides an opportunity to understand how

the members of a community change their behavior after that community is banned. Furthermore, knowledge of the drivers of inter-individual behavioral differences may permit moderators to monitor the post-ban activity of certain subsets of users more closely than others, which may lead to an increase in the efficacy of platform-wide moderation.

3.3 Methodology

As part of investigating whether different communities respond differently to a subreddit ban, we examine whether top users differ from random users in their change in activity and in-group language usage following community-level interventions. Specifically, we utilize natural language processing to track community activity after a subreddit ban, across 15 subreddits that were banned during the so-called "Great Ban" of 2020. We first identified words that had a higher prevalence in these subreddits than on Reddit as a whole prior to a ban. These words do not necessarily correspond to harmful speech but provide a linguistic signature of the community. The strengths and drawbacks of this approach are discussed in the discussion and appendix. We then compared the frequency of use of community-specific language, as well as the overall activity level of a user (i.e., the number of total comments), 60 days pre- and post-ban for (1) the 100 users that were most active in the banned subreddit 6 months prior to the ban and (2) 1000 randomly sampled non-top users. We predicted that top and random users that remained on the site following a subreddit ban would react differently to the ban, and we anticipated that there would be variation in how different communities responded to a ban.

3.3.1 Data Selection

We selected 15 subreddits banned in June 2020, after Reddit changed their content policies regarding communities that "incite violence or that promote hate based on identity or vulnerability" and subsequently banned approximately 2000 subreddits (i.e., "the Great Ban"). Based on a list of subreddits banned in the Great Ban [2] and an obscured list of subreddits ordered by daily active users [3], we examined the subreddits with more than 2000 active daily users and which had not previously become private subreddits. These most-visited subreddits were "obscured" by representing all letters except the first two as asterisks, but were de-anonymized as described in the appendix (section 3.8.1). By selecting highly active subreddits from the Great Ban we can compare many subreddits banned on the same date, and the differences in how their users responded. The list of subreddits we examined is included in section 3.3.5.

3.3.2 Data Collection

For each chosen subreddit, we collected all the submissions and comments made during the 182 days before it was banned. This is possible through the Pushshift API[4], which archives Reddit regularly, but may miss a minority of comments if they are deleted (by the author or by moderators) very shortly after they are posted [15]. We use this sample of the banned subreddits to identify users from the community and specific language used by the community. To accomplish the former, we examine the "author" field of each comment to get a list of users and how many comments

[2] https://www.reddit.com/r/reclassified/comments/fg3608/updated_list_of_all_known_banned_subreddits/

[3] https://www.redditstatic.com/banned-subreddits-june-2020.txt

[4] https://psaw.readthedocs.io/en/latest/

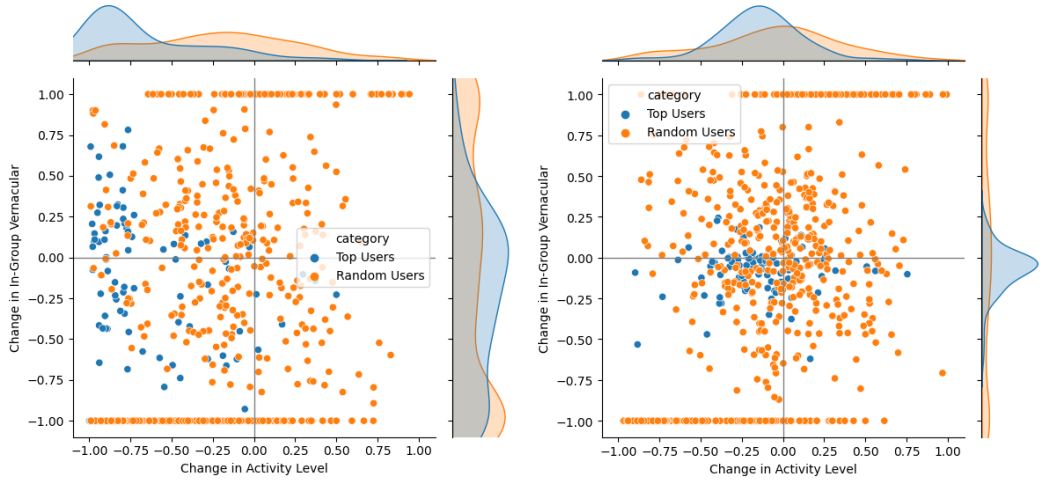
they made on the subreddit during the time frame prior to the ban.

To automatically determine in-group vocabulary words for a subreddit, we create a corpus of all text from the comments in a banned subreddit and compare it the baseline corpus to a corpus of 70 million non-bot comments from across all of Reddit during the same time frame. Bot detection is described in section 3.3.4. We can gather this cross-site sample by using comment IDs: every Reddit comment has a unique increasing numeric ID. By taking the comment ID of the first and last comments from our banned sample, and then uniformly sampling all comment IDs between that range and retrieving the associated comments, we can uniformly sample from Reddit as a whole over arbitrary time ranges.

We used this baseline corpus instead of a more standard English corpus because many such standard corpora rely on books, often in the public domain, whose language may be dated and more formal than Reddit comments. These corpora often also lack terms from current events such as sports team names or political figures, which occur frequently across large parts of Reddit.

3.3.3 Determining In-Group Vocabulary

We compare word frequencies between the two corpora to identify language that is more prominent in the banned subreddit than in the general sample. Since the two samples are from the same date range on the same platform, this methodology filters out current events and Reddit-specific vocabulary more than we would achieve by comparing to a general English-language corpus like LIWC [152]. Rather than comparing relative word occurrence frequency directly, which has pitfalls regarding low-frequency words that may only occur in one corpus, we apply Jensen-Shannon



(a) Ban effect on *r/gendercritical* users **(b)** Ban effect on *r/the_donald* users

Figure 3.1: Example plots comparing user behavior after a subreddit ban. Users from the top 100 and random samples are displayed in terms of their relative change in activity and change in in-group vocabulary usage. Distributions are displayed along each axis for convenience.

Divergence (JSD, see section 1.1.4 for more detail) which compares the word frequencies in the two corpora against a mixture text. JSD scores words highly if they appear disproportionately frequently in one corpus, even if they are common in both. For example, JSD identifies "female" as a top word in gender-discussion subreddits. Treating "female" as in-group vocabulary is undesirable for our specific use-case, where we would prefer to find language specific to the subreddit that is uncommon elsewhere. Therefore, we remove the top 10,000 most common words in the general corpus from both the general corpus and the subreddit corpus before processing. JSD functionality is provided by the Shifterator software package [51]. Based on the resulting JSD scores, we then select the top 100 words in the banned subreddit corpus, and treat this as our final list of in-group vocabulary. We used the top 100 words to maintain consistency with the distinctive vocabulary size used by Chandrasekharan et al. [27].

In the appendix, our approach is compared to the Sparse Additive Generative model (SAGE) of Chandrasekharan et al. [27] to show the additional flexibility of JSD as well as similarity of the results (see section 3.8.2).

3.3.4 Examining User Behavior

With a list of users from the banned community ranked by comment count and a list of in-group vocabulary, we are able to measure user behavior after the subreddit ban. Since larger subreddits can have tens of thousands to millions of users, we limit ourselves to examining two groups: (1) the 100 most active accounts from a banned subreddit, known as the "top users", and (2) a random sample of 1000 non-top users from the subreddit. In forming these lists of top and random users, we skip over accounts from a pre-defined list of automated Reddit bots as well as users that have deleted their accounts and cannot have their post histories retrieved. Additionally, as our focus for this study is users who used in-group language and who continue to use the platform, we omit users that have never used in-group vocabulary pre- or post-ban or who have zero comments post-ban. All forms of user-filtering are discussed further in the appendix (section 3.8.4).

For each user, we download all the comments they made in the 60 days before and after the subreddit ban. We compare the number of comments made before and after the ban to establish a change of activity, on a scale from -1 to 1, with -1 indicating "100% of the user's comments were made prior to the ban", 0 indicating "an equal number of comments were made before and after the ban", and 1 indicating that all of their comments on Reddit were made after the ban. We can similarly track the user's use of in-group vocabulary on a scale from -1 to 1, for "100% of their in-group

vocabulary usage was before the ban" to "all uses of in-group vocabulary were post-ban". This is calculated as the fraction of posted words that were in-group vocabulary after the ban, minus the fraction of posted words that we in-group vocabulary before the ban, divided by the sum of the fractions.

Difference between vocabulary usage after and before ban

$$\frac{r_a - r_b}{r_a + r_b} \tag{3.1}$$

Total vocabulary used before and after ban

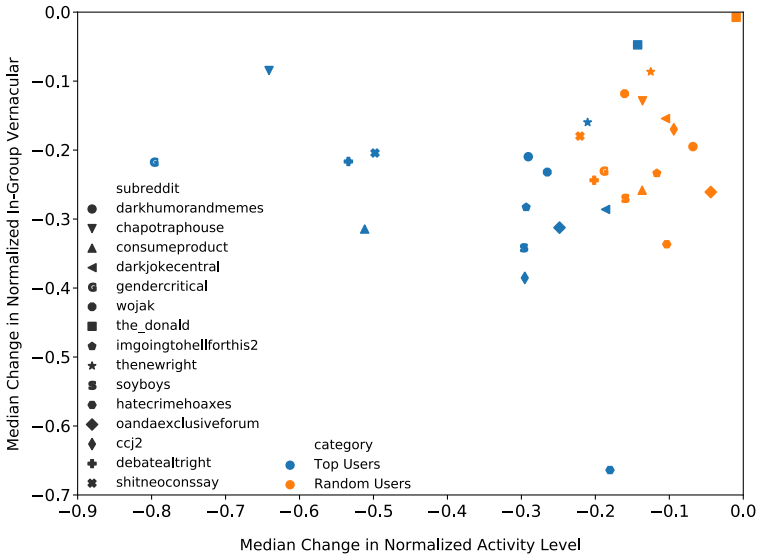
Examples of results for individual subreddits are shown in Fig.3.1.

3.3.5 Statistical Methods

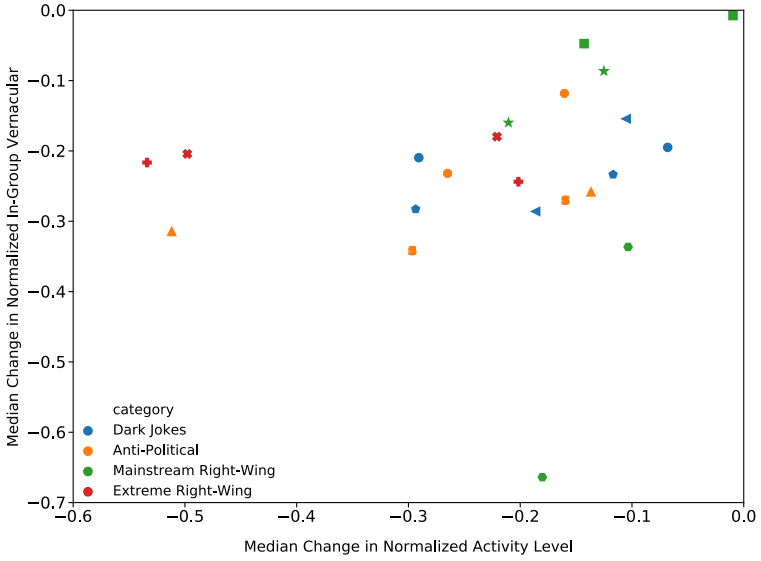
Category	Subreddits
Dark Jokes	darkjokecentral, darkhumorandmemes, imgoingtohellforthis2
Anti-Political	consumeproduct, soyboys, wojak
Mainstream Right Wing	the_donald, thenewright, hatecrimehoaxes
Extreme Right Wing	debatealtright, shitneoconssay
Uncategorized	ccj2, chapotraphouse, gendercritical, oandaexclusiveforum

Table 3.1: Subreddit categorization by qualitative assessment of content

We do not necessarily expect all subreddits to respond to a ban in the same way. From the user data for the 60 days before and after the subreddit's banning, we examined whether there was any difference between subreddits for (1) the proportion of a user's total posts that occurred postban vs preban and (2) the proportion of a user's total in-group vocabulary that occurred postban vs preban. We also explored whether a user's engagement in a subreddit (i.e., whether they were a top or random user) influenced either measure. To examine the predictors of the proportion of a



(a) Comparison of top/random users in all 15 subreddits



(b) Comparison of top/random users across by categories

Figure 3.2: Comparison of top and random user behavior changes across fifteen subreddits banned after a change in Reddit content policy in January, 2020. (a) Top users show more significant drop-offs in posting activity after a ban, but have around the same change in in-group vocabulary usage as a uniform sampling of subreddit participants. (b) Ban impact on eleven subreddits categorized by content. Each subreddit appears twice, representing top and random users. Four uncategorized subreddits are excluded from the plot. Trends are summarized in section 3.4.

user's total posts that occurred postban vs preban, we ran a generalized linear mixed model with a binomial error distribution. This model included the ratio of a user's posts after the ban to their posts before the ban as the predictor, and subreddit identity and user engagement (i.e., top or random) as fixed effects. To examine the predictors of pre-ban vs post-ban total in-group vocabulary, we ran a second generalized linear mixed model with a binomial error distribution. Its predictor was the ratio of the number of in-group vocabulary words a user used after the ban to the number of in-group vocabulary words that they used before the ban. Subreddit identity and user engagement (i.e., top or random) were fixed effects. For both models, we included user identity (i.e. top or random) as a random effect, since some users were active in more than one of the studied subreddits. Additionally, we used a likelihood ratio test (LRT) to explore whether there was an overall effect of subreddit identity on the proportion of a user's total posts that occurred postban vs preban, and the proportion of a user's total in-group vocabulary that occurred postban vs preban. In the LRT, we compared each described model to a model without subreddit identity. We also used LRTs to compare models with and without user engagement to assess whether there was an overall effect of user engagement on either measure.

We performed statistical comparisons in order to understand whether users' vocabulary and activity differed before and after the ban, as well as whether top and random users of a given subreddit experienced similar shifts.

To confirm the shifts displayed in fig. 3.2a are meaningful we performed Wilcoxon Signed-Rank tests ($\alpha = FDR = 0.05$) on the normalized vocabulary ratios and normalized activity ratios before and after the ban. Except for users of the_donald (both user types) and the top users of chapotraphouse, these tests decreases in-group

vocabulary usage in all subreddit/user-type pairs. The same tests showed the ban had a significant effect on all subreddit/user-type pairs in terms of activity level except for the random users of the_donald, though these effects were not all decreases.

We used the Wilcoxon rank sum test to compare the previously defined metrics for vocabulary shift and activity shift between the top and random users within each subreddit. The p-values for each individual comparison at the subreddit level were corrected using false discovery rate (FDR), and are illustrated in fig. 3.3.

3.3.6 Subreddit Categorization

To better understand our results, we categorized each banned subreddit as "dark jokes", "anti-political", "mainstream right wing", and "extreme right wing", as shown in section 3.3.5. These categories encompass eleven of our fifteen subreddits, leaving four that are significantly distinct from their peers. Note that the "uncategorized" subreddits are not necessarily difficult to classify (for example, r/gendercritical is a trans-exclusionary radical feminist subreddit), but without similar banned subreddits of comparable size we cannot suggest that results for these subreddits are generalizable. While these categories were chosen based on qualitative assessment of each subreddit's content, they are verified by a quantitative comparison of the unique vocabulary of each subreddit available in the appendix.

3.4 Results

By comparing the median change in activity and vocabulary usage among top and random users, we found a consistent pattern: Top users, for every subreddit studied,

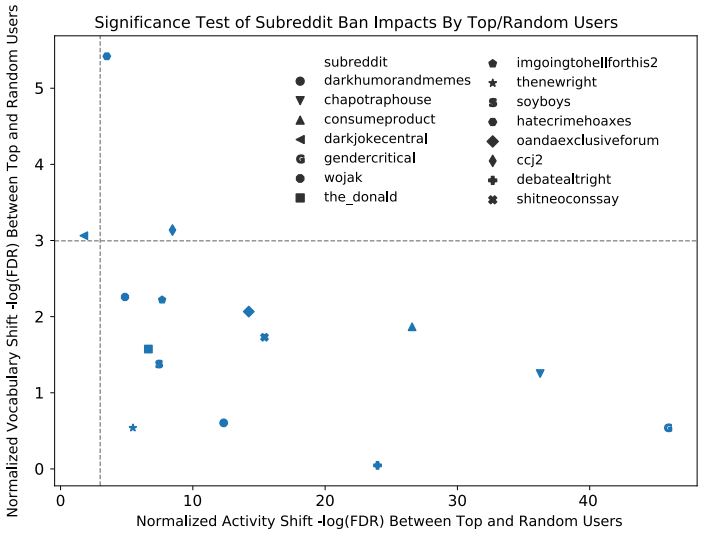


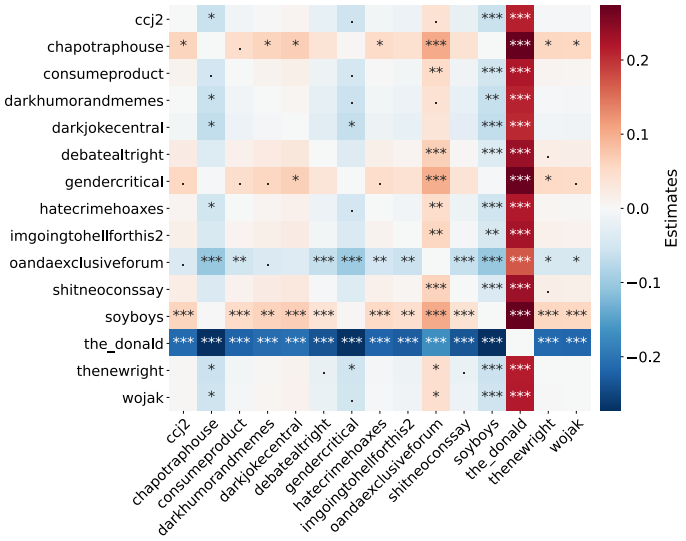
Figure 3.3: Scatterplot showing differences in activity and vocabulary shifts between top and random users of each subreddit. Each axis shows the statistical significance, expressed as -log(FDR), of either activity (x-axis) or vocabulary (y-axis) shift. Dashed lines indicate significance at a threshold of 0.05, such that subreddits with greater values show significant differences between top and random users.

decrease their activity more than their peers. This result is important to keep in mind when a uniform sampling of subreddit users post-ban may indicate that a community ban was ineffective. We do not find as consistent a difference between top and random user when looking at vocabulary change; suggesting that while bans may drive harmful users to inactivity, they are less clearly effectual at reforming user behavior. These results are summarized in fig. 3.2a.

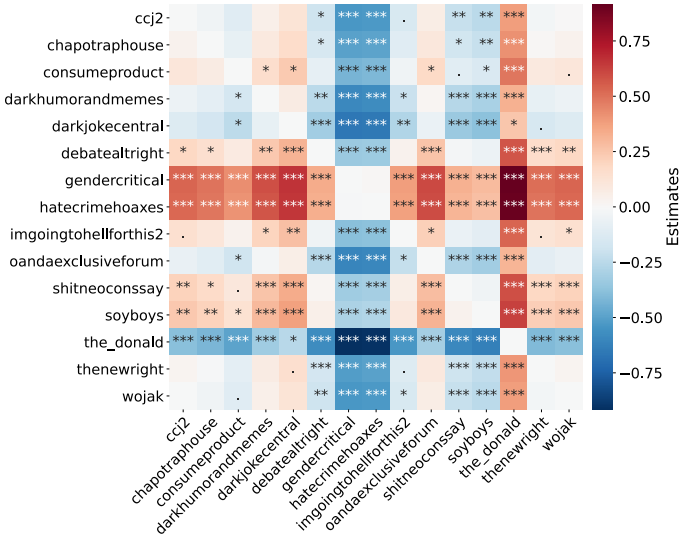
To confirm our findings, we tested the statistical significance of differences between top and random distributions for each subreddit, illustrated in fig. 3.3. In all subreddits, there was a significant difference between top and random user changes in either activity shifts, vocabulary shifts, or both. Considering a significance threshold on the false discovery rate, FDR < 0.05, we found two subreddits (r/ccj2 and r/hatecrimehoaxes) that show significant differences in both shifts. The subreddit

r/darkjokecentral shows significant differences between top and random users in vocabulary shift, but not activity; whereas the rest of the subreddits show differences in activity but not vocabulary shift between top and random users.

We found that, controlling for user engagement (i.e., whether a user was a top or random user), there was a significant overall effect of subreddit identity on both the proportion of a user's total posts that occurred postban vs preban (LRT, Chi-squared = 133.730, $p < 0.001$) and the proportion of a user's total in-group vocabulary that occurred postban vs preban (LRT, Chi-squared = 239.680, $p < 0.001$). Controlling for subreddit identity, there was also a significant overall effect of user engagement on the proportion of a user's total posts that occurred postban vs preban (LRT, Chi-squared = 23.452, $p < 0.001$) and the proportion of a user's total in-group vocabulary that occurred postban (LRT, Chi-squared = 220.020, $p < 0.001$). Postban posts made up a lower proportion of a user's total posts and postban use of in-group vocabulary made up a lower portion of a user's total in-group vocabulary use for top users compared to random users (fig. 3.4). There were a few subreddits that were significantly different from most or all of the other subreddits. For example, in r/the_donald, postban posts comprised a higher proportion of a user's total posts, compared to all other subreddits (fig. 3.4a), and postban use of in-group vocabulary comprised a higher portion of a user's total in-group vocabulary use, compared to all other subreddits (fig. 3.4b). Postban posts also comprised a higher proportion of a user's total posts in r/oandaexclusiveforum, compared to most other subreddits, while postban posts comprised a lower proportion of a user's total posts in r/soyboys, compared to most other subreddits (fig. 3.4a). The proportion of a user's total in-group vocabulary that occurred postban was lower for both r/gendercritical and r/hatecrimehoaxes,



(a) Proportion of Total Posts made Post / Pre-ban



(b) Proportion of Total In-Group Vocabulary used Post / Pre-ban

Figure 3.4: Visualization of GLMM results showing differences between subreddits in postban behavior. For each row, blue cells indicate that the subreddit in a given column had a lower proportion of postban activity/ingroup vocabulary use than the subreddit in that row, while red cells indicate that the subreddit in a given column had a higher proportion of postban activity/ingroup vocabulary use than the subreddit in that row. · indicates p < 0.10. * indicates p < 0.05. ** indicates p < 0.01. *** indicates p < 0.001.

compared to most other subreddits (fig. 3.4b).

Category	Activity Impact	Vocabulary Impact
Dark Jokes	Minimal	Minimal
Anti-Political	Top users less active	Decrease among top users
Mainstream Right Wing	Minimal	Inconsistent
Extreme Right Wing	All users decrease, especially top users	Minimal

Table 3.2: The impact of subreddit bans within each category.

3.5 Discussion

Past work has been quick to conclude that subreddit bans either are [27, 140, 155] or are not [64] effective at changing user behavior. We have found that results differ between subreddits and between more and less active users within a subreddit. Since many prior studies on banning efficacy focus on one to two subreddit case studies, these distinctions may not have been apparent in some previous datasets.

To automatically study a larger number of communities, we tackle the simpler problem of tracking user activity and use of in-group language rather than more subjective harmful language. This approach has strengths and drawbacks. On the one hand, in-group language is easier to automatically identify with little expert knowledge or human intervention, while also including lesser known slang terms or dog whistles that could be harmful. On the other hand, our approach requires a large reference corpus that controls for relevant features of the studied corpus to produce meaningful results. For Reddit, using non-banned subreddits as a baseline corpus allows us to automatically study changes in activity and language around community bans while requiring little expert knowledge on these communities. However, choosing

a reference corpus may be more challenging on other platforms without a broader "mainstream" population (such as alt-tech platforms), with small populations, or without a clear means of sampling the overall population (such as Slack, Discord, and Telegram).

Our study examines 15 subreddits with over 5000 daily users that were banned simultaneously after a change in Reddit content policy, and our results suggest that subreddit bans impact top and random users differently (in agreement with prior studies such as Hazel Kwon and Shao [68], Ribeiro et al. [137], and Thomas et al. [155]) and that community-level banning has a heterogeneous impact across subreddits.

Additionally, we see patterns in subreddit responses to bans that loosely correlate with the type of content the community focused on, summarized in section 3.4 and illustrated in fig. 3.2b. Dark joke subreddits were banned for casual racism, sexism, or other bigotry, do not have as clearly defined in-group language, and were largely unaffected by bans. Users are not more or less active, and use similar language pre and post-ban. Anti-political subreddits, who ridicule most activism and view social progressiveness as performative, were moderately impacted by bans. Top users from these communities became less active after the ban, and randomly sampled users commented using less in-group language. Mainstream right-wing communities show the least consistency in ban response. The most impacted subreddits were extreme political communities that blatantly advocated for white supremacy, anti-multiculturalism, and fascism. These communities saw median top user activity drop to under a third of pre-ban levels, followed by a significant decrease in random user activity, and a modest decrease in in-group vocabulary usage (about -0.2 to -0.3 for all user groups). Since our sample includes only two to four subreddits per category,

these trends are not robust but suggest that some pattern might exist within the heterogeneous responses to community-level bans. These results could guide future moderation of online spaces and therefore merit further investigation.

3.6 Conclusion

We have provided a broad investigation of the impact of banning online communities on the activity and in-group vocabulary of the users therein. Our work expands the scope of other studies on this subject, both in terms of the number and types of communities examined. Through this more comprehensive analysis, we have demonstrated heterogeneity in the impact of bans, depending on the type of subreddit and the level of user engagement. We found that top users generally showed greater reductions in activity and in-group vocabulary usage, compared to random users. We also found that the efficacy of banning differs across subreddits, with subreddit content potentially underlying these differences. However, while we provide strong evidence of heterogeneity in ban efficacy, even more comprehensive research must be conducted on a larger group of subreddits in order to fully understand the dynamics behind this heterogeneity.

3.7 Future Work

This study finds heterogeneity in the outcomes of the largest online communities banned on Reddit at the community level and at the individual level. Though we find a clear trend relating outcomes to pre-ban activity level between the top and

random users, there are likely other factors at play. Future work could investigate which factors correlate with individual user responses to subreddit bans, including: user demographics (both those directly measurable, such as age of account, and those like gender or country of residence ascertained via tools such as machine learning classifiers), more complex activity metrics (e.g. position of users in interaction networks within the community), and activity in other communities (as measured by number and label of other communities engaged with and level and response of engagement within those communities).

While we find evidence that community-level responses to bans loosely correlate with the content of the subreddit, our limited sample size of 15 subreddits precludes any thorough quantitative comparisons. Unfortunately, including subreddits with fewer users than the 15 we selected would make community-level statistics less consistent. Were a future study to include large banned subreddits from before or after the "great ban", identifying the factors and mechanisms that contribute to the differences in subreddit responses would be an important contribution. Potential such factors include: the demographic makeup of the communities, interaction types within the community (potentially measured via network analysis of the comment interaction network of the community), and position in a subreddit-subreddit network of shared users. Studies examining longer-term impacts of community bans would also benefit from considering when some communities attempt to "rebuild" in a new subreddit, versus integrate into existing subreddits, or rebuild off Reddit entirely.

However, we believe the most valuable insights may come from embracing more holistic, qualitative methodologies to characterize these banned communities and their responses to moderation. While quantitative metrics indicate heterogeneous commu-

nity responses, researchers from anthropology and sociology, as well as communications and media studies, may find additional depth in community and user response to censorship. Computational linguists may be able to refine techniques for detecting in-group vocabulary, while linguists and cultural evolution specialists may be best equipped to determine how these vocabularies drift over time. Finally, social computing experts may be in the best position to adapt these multidisciplinary findings to improve platform moderation tools and policies.

3.8 Appendix

3.8.1 Banned Subreddit De-Obfuscation Process

We used a report of the subreddits banned in the "Great Ban" ranked by daily average users (DAU) [5]. The top 20 subreddits with the highest DAU were reported with their names in clear text. The rest of the subreddits had their names obscured, showing only the first two letters and the remaining characters replaced by asterisks.

To de-obfuscate these, we used the subreddit *r/reclassified* [6], in which users report banned and quarantined subreddits. We used the Pushshift API to recover posts for the week after the "Great Ban", and selected those that had been flagged with the flair *BANNED*.

We then used the following routine to identify the obfuscated banned subreddits from the first list:

For a given sequence of two initial letters and a given subreddit name length, let

[5] https://www.redditstatic.com/banned-subreddits-june-2020.txt
[6] https://www.reddit.com/r/reclassified/

N be the number of obscured subreddits with this sequence and name length. Let M be the number of purged subreddits with this initial sequence of letters and length. The M purged subreddits are therefore candidates for the N obscured subreddits.

If $N \geq M$, disambiguate the N obscured subreddits as the M purged subreddits. Any unmatched obscured subreddits are omitted from our analysis.

If $N < M$, manually select the N most-populous subreddits from the M candidate subreddits. Number of commenters was manually researched in the https://reddit.guide/ page for the candidate subreddits.

3.8.2 Comparison of Keyword-Selection Methods

The identification of community specific keywords or the identification of hateful speech is an essential part of the pipeline for any kind of analysis on the effect of interventions on online speech. Just as there are numerous methods for the identification of hateful speech [53, 123, 145, 99], there are numerous related methods for the identification of community-specific keywords. Chandrasekharan et al. [27] used a topic modelling framework to identify keywords for their study called the Sparse Additive Generative model (SAGE) which compares "... the parameters of two logistically-parameterized multinomial models, using a self-tuned regularization parameter to control the tradeoff between frequent and rare terms." The core of this method, the parameter comparison of two logistically-parameterized multinomial models, performs a similar task as our ranking of the contributions of each term to the overall Jensen Shannon Divergence (JSD), and the regularization parameter

performs a similar task as our explicit removal of the most common terms in our baseline corpus. As both our methodology and that of Chandrasekharan et al. [27] perform comparable steps to achieve a comparable outcome, one would expect comparable results. This is somewhat the case when the results are defined for both methods as we can see in the table 3.4 below by considering the intersection of terms. However, an important feature of Jensen Shannon Divergence is how it addresses the "out-of-vocabulary problem" where an instance of a term of any frequency in one corpus has infinitely higher relative frequency than in a compared corpus if that compared corpus does not contain that term. Simplistically, JSD addresses this issue by comparing both corpora to a reference corpus made up of an amalgamation of the two. The SAGE methodology on the other hand, does not have an answer to this problem laid out and so without additional modifications, the SAGE coefficients for such terms that appear in a subreddit of interest but not in a baseline corpus are undefined, and a list of keywords is methodologically impossible to ascertain. As such, we argue that using our JSD-based methodology is more robust to this out-of-vocabulary problem and thus more widely applicable in a variety of settings. Additionally, we view the explicitness of our keyword selection methodology as an advantage compared to the relative "black box" nature of SAGE.

However, despite the fact that the SAGE-based keyword selection methodology yielded undefined values for a number of the subreddits we studied, given the importance of Chandrasekharan et al. [27] as foundational to our work, we developed a small extension to the SAGE-based methodology which provides estimates of what the SAGE coefficients would be with a baseline corpus of the entire population of Reddit comments rather than only a sample (note that such a baseline corpus would

no longer face this out-of-vocabulary problem as all terms in the subreddit of interest would appear in the population since the subreddit of interest is part of the population). The way these estimates were reached was to use additional known metadata to estimate the counts of all the terms in the baseline corpus as well as the terms in the subreddit of interest which did not appear in the baseline. This was achieved as follows: First, take the frequency counts of each word in the baseline corpus and normalize them to calculate the empirically estimated probability mass function for words in the population of all comments on Reddit for our 6 month timeframe. Second, estimate the number of words on Reddit during this timeframe by taking the exact number of comments on Reddit during this timeframe (calculated by subtracting the first comment ID from this timeframe from the last comment ID from this timeframe) and multiplying this number by the mean number of words per comment in the baseline corpus of 70 million random comments. Third, multiply this estimated number of words on Reddit by the estimated probability mass function for each word to calculate the estimated count of each word in the population rather than the sample. Fourth, add the counts of the out-of-vocabulary terms to these estimated population-sized counts. In the event that those terms appeared only in the subreddit of interest and nowhere else on Reddit during the timeframe examined, this count will be the exact count for that term in the population and it will be at the approximate relative scale when compared to the estimated counts of the other terms in this new estimated population corpus. Using this newly estimated "population" baseline corpus, we follow the SAGE-based methodology as in Chandrasekharan et al. [27] to determine the set of keywords identified by this methodology. Note that in the event that there are no out-of-vocabulary terms, this method simply scales up the

frequencies by a constant amount for each term and as a result, reduces exactly to if this extra step had not been performed, but for cases where the out-of-vocabulary problem presents itself, this allows us to gather a list of terms comparable to that methodology.

Subreddit	**Intersection**
ccj2	20
chapotraphouse	51
consumeproduct	61
darkhumorandmemes	46
darkjokecentral	17
debatealtright	35
gendercritical	53
hatecrimehoaxes	33
imgoingtohellforthis2	36
oandaexclusiveforum	9
shitneoconssay	31
soyboys	51
the_donald	56
thenewright	57
wojak	34
MEAN	**39.65**

Table 3.3: Number of shared vocabulary words between our JSD-based keyword selection methodology and the SAGE-based methodology

Examining figure 3.5, we first notice that for the most part, most subreddit/user-type pairs are in relatively similar positions under the SAGE methodology as under the JSD-based keyword selection, especially when compared relative to each other. Chandrasekharan et al. [27] found strong negative shifts in in-group vocabulary usage after bans. Upon reproduction of their methodology, we also find stronger negative shifts, including several subreddit/user-type pairs which exhibit a median value of the maximum possible negative vocabulary shift (-1). I.e. the majority of users in these subreddits used at least one SAGE-selected keyword prior to the ban and none

thereafter. Examining the data directly, we find that among the subreddit/user-type pairs where this occurred, all five had over half of their users use a SAGE-identified in-group vocabulary word between one and three times only prior to the ban. Additionally, three out of five had a majority use a SAGE-identified in-group vocabulary word one to three times prior to the ban and then zero times after the ban. Under the JSD-based methodology, no subreddit/user-type exhibited behavior where the majority of the users ceased all vocabulary usage after the ban.

The implication that the words chosen by SAGE are not used frequently by a majority of the users of subreddits they are selected from, and are thus not ideally representative, is further supported by the fact that a much larger portion users initially collected had to be omitted due to having zero vocabulary word usage before or after the ban. For the JSD-based methodology, an average of 263 of the initially collected 1000 users were omitted for having never used a single JSD-selected keyword at any time. Under the SAGE-based methodology, this number was 158 users higher on average. I.e. there was a substantially greater portion of users who used no SAGE identified vocabulary words either before or after the ban than users who used no JSD-identified vocabulary words.

The omissions mentioned above are the only cause of differences in activity shift between the the two methodologies. Apart from which users were omitted, the users studied under each methodology were identical and thus had identical activity shifts.

Subreddit	1st Match	2nd Match	3rd Match
ccj2	imgoingtohellforthis2 (4)	darkhumorandmemes (3)	chapotraphouse (2)
chapotraphouse	shitneoconssay (8)	consumeproduct (7)	thenewright (5)
consumeproduct	wojak (37)	soyboys (37)	shitneoconssay (19)
darkhumorandmemes	imgoingtohellforthis2 (22)	darkjokecentral (18)	wojak (11)
darkjokecentral	darkhumorandmemes (18)	imgoingtohellforthis2 (7)	wojak (4)
debatealtright	shitneoconssay (49)	thenewright (30)	consumeproduct (14)
gendercritical	darkhumorandmemes (5)	consumeproduct (3)	soyboys (2)
hatecrimehoaxes	imgoingtohellforthis2 (14)	thenewright (6)	debatealtright (5)
imgoingtohellforthis2	darkhumorandmemes (22)	thenewright (16)	soyboys (14)
oandaexclusiveforum	darkhumorandmemes (4)	wojak (4)	imgoingtohellforthis2 (3)
shitneoconssay	debatealtright (49)	thenewright (29)	consumeproduct (19)
soyboys	consumeproduct (37)	wojak (26)	imgoingtohellforthis2 (14)
the_donald	thenewright (15)	shitneoconssay (11)	consumeproduct (7)
thenewright	debatealtright (30)	shitneoconssay (29)	imgoingtohellforthis2 (16)
wojak	consumeproduct (37)	soyboys (26)	imgoingtohellforthis2 (13)

Table 3.4: Comparison of subreddits based on number of shared terms in their respective top 100 in-group vocabulary. These number of shared terms, shown in parenthesis, reinforce qualitative categorization in section 3.3.5

3.8.3 Validation of Subreddit Categories by Vocabulary Overlap

We initially classified each subreddit by a qualitative assessment of community content. However, we can hypothesize that subreddits with similar focuses are more likely to share in-group vocabulary terms, or conversely, that unrelated subreddits with divergent content are unlikely to share in-group vocabulary. Therefore, if our categorization is accurate, subreddits in each category should share more in-group vocabulary with one another than with other subreddits. This is easily tested, and the results are shown in table 3.4.

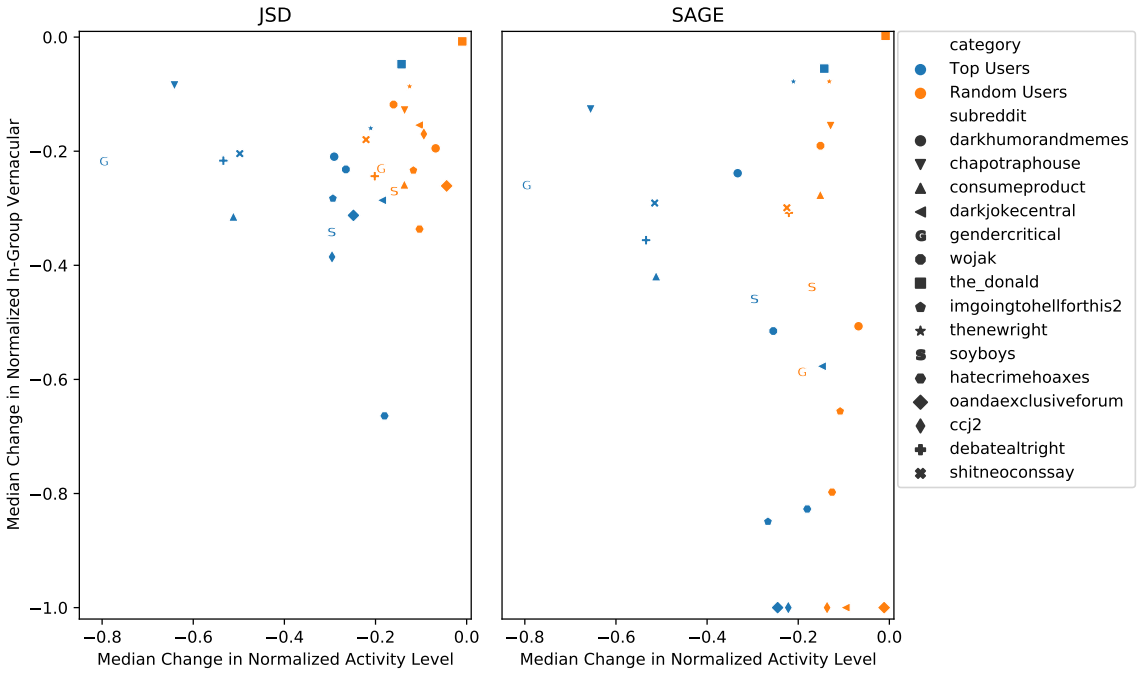


Figure 3.5: Comparison of top and random user behavior changes under different keyword selection methodology. The subplot on the left corresponds to 3.2a in the main text. The differences in activity shift between the two plots are minute and only due to omission of slightly different users for having no in-group vocabulary usage before or after the ban. The relative positions on the vocabulary shift axis remain largely the same except for a wider distribution and several subreddit user-type pairs exhibiting the maximum possible negative shift as the median.

3.8.4 Accounts Omitted from Analysis

In order to limit the analysis to human users and exclude any unobservable or misleading data, we excluded from all parts of the pipeline of this research (from keyword identification to vocabulary shift analysis) any comment which was made by a username in an amassed list of non-human 'bot' users. Additionally, we excluded any comment which was made by a user who deleted their account between the time of posting and the time of data ingestion by PushShift, as comments made by these

users all present with the indistinguishable username "[deleted]." We used a list of bots curated by botrank.pastimes.eu, which itself uses its own Reddit bot to scrape comments searching for replies to accounts indicating that the replying user considers the account to be a bot. These comments are a common practice on Reddit and take the form of users indicating their approval or disapproval of an account they perceive to be a bot via the phrases "Good bot/good bot" and "Bad bot/bad bot" respectively. The system that populates botrank.pastimes.eu scrapes from all comments on Reddit at intervals and compiles a list of accounts who have had either "good bot" or "bad bot" replied to them, as well as the number of times this has been done for each such account. The higher the sum of the counts of "good bot" and "bad bot" replies, the more users who have identified the given account as a bot (and are expressing their approval or disapproval of this account). Thus, accounts which have high counts of these replies can be considered as very likely to be bots. As such, we assembled the majority of the list of accounts we excluded from our analysis via identifying each such account in the above mentioned compilation which had over 300 occurrences of users reply either "good bot" or "bad bot" to them. This contributed 263 accounts we excluded. Additionally, we manually identified two other accounts below this threshold of 300 occurrences as bots by combing through the data ('dark-repostbot', and 'tweettranscriberbot'). With the addition of the '[deleted]' accounts, this resulted in a total of 266 usernames for which comments were excluded from our analysis, which are included in supplementary material.

Because the focus of our study was users who continued to use the platform and who used in-group language, we omitted users who had zero comments after the ban and users who had zero instances of in-group vocabulary usage before or

after the ban. No top users fell into either of these categories as they all used in group language either before or after the ban and all made at least one comment after the ban. The breakdown of how many users this final sequence of omissions results in amongst the random users, broken down as subreddit:(number users omitted for having zero postban comments, number users omitted for having no in-group vocabulary usage), is as follows: oandaexclusiveforum: (171, 239); ccj2: (174, 264); darkjokecentral: (132, 468); darkhumorandmemes: (146, 477); shitneoconssay:(223, 119) ; imgoingtohellforthis2:(141, 358); consumeproduct:(147, 292); the_donald:(94, 332); debatealtright:(257, 118); gendercritical: (207, 278); chapotraphouse:(108, 222); soyboys:(203 , 214); hatecrimehoaxes:(141, 113); thenewright:(128, 190); wojak:(137, 257).

Chapter 4

Measuring Centralization of Online Platforms Through Size and Interconnection of Communities

Foreword

There are several notions of "control" on online platforms. Platform administrators set social content policies for what is permissible on their site, with the ability to ban users or content, and they control what interactions between users are possible, through adding or removing functionality to the platform. Administrators' abilities are to some extent vetoed by the owners of technical infrastructure, including hosting providers and content distribution networks; if administrators permit widely objectionable content that does not cross the line of illegality, infrastructure providers may choose to terminate service agreements and take the platform offline, as seen with

8Chan [73], Kiwi Farms [104], Parler [77], and others. Finally, influential users and community organizers within a platform exert social control by choosing what content to post and to help proliferate.

Rich-get-richer dynamics are common on social media [14]. On user-centric platforms like Twitter, popular accounts with many followers are more "visible" in that their posts are re-posted further along a social graph, receive more engagement, and are more likely to be promoted by engagement-maximizing recommendation algorithms, ultimately bringing the popular accounts more followers and engagement. On community-centric platforms like Reddit, popular subreddits have more users, posts, and comments, and so have more content to engage with than smaller peers, which draws in new users, further fueling community growth. Even in contexts like open-source software development, more active projects with many contributors are more likely to attract new contributors than smaller more obscure projects (see chapter 2). This combination of growth over time, and newcomer preferential attachment (that is, a tendency for new users to engage more with popular users and communities than with unpopular ones) makes highly influential hubs appear to be a natural function of social media.

A tendency towards centralization amplifies power dynamics. Users with social or moderation control of influential hubs exert immense influence over a social platform. Platform administrators can enforce social policies through exerting control over hubs, as seen in chapter 3, where Reddit chose to eliminate unwanted communities rather than policing behavior site-wide.

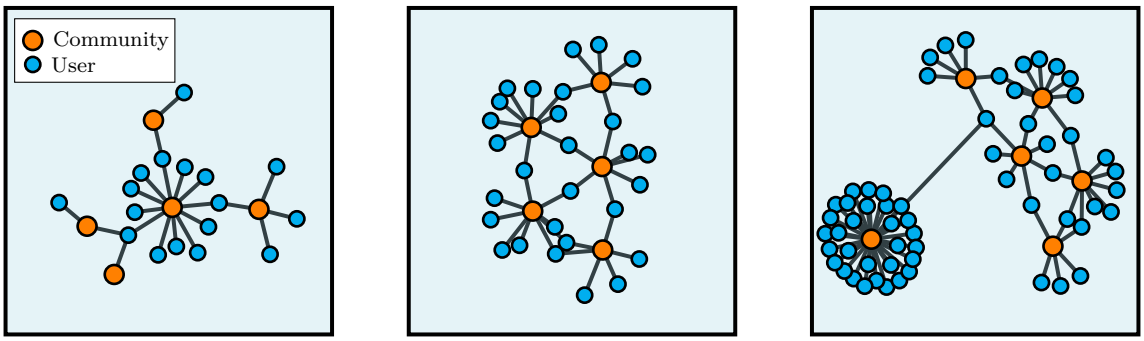
Several platforms are designed to subvert social control of major sites. For example, alt-tech platforms including BitChute [158] and Voat [110], largely mirror

mainstream counterparts YouTube and Reddit, but permit content deplatformed from their more conventional peers. These sites are still centrally governed by platform administrators and hosting providers, and differentiate themselves through the policies those governors choose to enforce. By contrast, Federated platforms like Mastodon attempt to eliminate centralized governance altogether; communities are hosted across many interoperable servers, with independent administrators, moderators, hosting providers, and social policies. This limits the ability for any small party to exert influence over the entire platform, and allows different community spaces to have divergent norms and standards, negating the challenges of enforcing a single moderation standard across a widely diverse public sphere [170]. However, even in federated services, preferential attachment dynamics lead some servers to grow much larger than others, and server administrators can exert indirect control over one another by threatening to withhold interoperability, effectively isolating a community from the rest of the platform until it is compelled to adopt shared community standards.

In this chapter I examine how platforms differ in social dynamics, through the lens of community-size and interconnection. I am interested in the community-size distribution, that is, to what degree rich-get-richer dynamics have led to a small minority of communities containing the majority of users and content on a platform. Equally important is the level of interconnection, or inversely, insularity, among communities. If a large community has few connections to the rest of the platform then it creates a fiefdom, where moderators and administrators may have substantial influence within their walls but little influence beyond. By contrast, if a community is both large and shares many users with other communities, then policy decisions made within its borders may influence the information spread to many other communities.

Abstract

Decentralization of online social platforms offers a variety of potential benefits, including divesting of moderator and administrator authority among a wider population, allowing a variety of communities with differing social standards to coexist, and making the platform more resilient to technical or social attack. However, a platform offering a decentralized architecture does not guarantee that users will use it in a decentralized way, and measuring the centralization of socio-technical networks is not an easy task. In this paper we introduce a method of characterizing inter-community influence, to measure the impact that removing a community would have on the remainder of a platform. Our approach provides a careful definition of "centralization" appropriate in bipartite user-community socio-technical networks, and demonstrates the inadequacy of more trivial methods for interrogating centralization such as examining the distribution of community sizes. We use this method to compare the structure of five socio-technical platforms, and find that even decentralized platforms like Mastodon are far more centralized than any synthetic networks used for comparison. We discuss how this method can be used to identify when a platform is more centralized than it initially appears, either through inherent social pressure like assortative preferential attachment, or through astroturfing by platform administrators, and how this knowledge can inform platform governance and user trust.



(a) Centralized **(b)** Decentralized **(c)** Ambiguous

Figure 4.1: The influence of a community is tied to both its size and topological role in a network. In the centralized network, the orange community at the center both has the largest population of blue users, and serves as a bridge between four other communities. In the decentralized example, communities are of variable size, but none have a pivotal position to influence their peers. In the ambiguous case, one community is much larger, but the remaining network matches the "decentralized" example. Neither a distribution of community sizes nor purely structural measurements like betweenness centrality or graph conductance adequately capture this notion of community-level influence.

Online social spaces are vulnerable to centralized authorities making decisions that negatively affect the community. In 2022, the Software Freedom Conservancy recommended that all developers migrate their projects away from GitHub [56], after Microsoft bought the software development collaboration platform and used open source projects as training data for their commercial CoPilot software, in violation of open source licenses and community standards. The same year, users and advertisers departed Twitter after its purchase by Elon Musk and subsequent changes in community policy and staffing, including firing content moderators [93] and reinstating a number of accounts banned for violating the platform's hateful content and harassment policies [72]. Reddit moderators have historically engaged in blackouts to protest administrative policies [80], and these trends are ongoing; in June, 2023, Reddit announced plans to begin charging for API access, sparking warnings from

scientists [39], outrage among users, and a protest across nearly 9,000 subreddits, the long-term effects of which remain to be seen. As users express dissatisfaction with platform administrators, they have sought alternative platforms without centralized control, leading to the rapid growth of "federated" platforms like Mastodon [176] and Bluesky[1]. Alternatively, other users have promoted self-hosted platforms, such as independently operated git servers, or peer-to-peer hosting solutions such as the Interplanetary File System (IPFS) or web-torrent video hosting software PeerTube. Some deplatformed users have also responded by creating close facsimiles of existing centralized platforms with extremely permissive content-policies, frequently called "alt-tech" platforms [42].

What exactly is "centralization" in an online social network? Does it describe ownership of the platform? Its technical infrastructure? The creation and enforcement of community norms? The distribution of activity and reach of content producers? Centralization has long been ill-defined by academics [47], and "decentralization" joins as a widely-used but contextually redefined term today [40]. Of particular interest to us is a notion of group social influence: How much does one community impact others across a platform? For example, how independent are subreddits on Reddit, and how closely interlinked are Mastodon instances, the nascent "decentralized Twitter alternative?" Our goal is to measure the influence of a socio-technical platform's subcommunities on their peers, providing a mesoscale metric to quantify centralization at an inter-group level.

Measuring group-level influence has applications in content moderation, platform governance, and public awareness of administrative behavior. First, it allows admin-

[1]Bluesky is still in beta, and while the protocol is federated, only one instance exists at the time of writing.

istrators, moderators, and community organizers to identify and proactively avoid risks to community welfare. For example, if Mastodon's goal is to create a decentralized Fediverse then measuring the influence of a large instance can inform decisions on when to close registration on that instance, or stop recommending it on new-user onboarding websites like instances.social, to direct new users to more diverse instances. If administrators of smaller instances want to mitigate the viral spread of information from influential instances, they can de-emphasize posts from the larger instance, for example by hiding them from the federated feed. This aligns with recent proposals to design social media for abusability, by making design choices that limit usability for a minority of users to protect usability for the majority [17].

Next, measuring community influence allows platform users to identify when administrators are engaging in "decentralization astroturfing." Some administrators misrepresent the level of decentralization or community self-governance on their platforms, allowing them to abdicate responsibility for community moderation and social policy. For example, Bluesky has no dedicated moderation or trust and safety team, because they publicly aspire to provide tools and protocols for communities to self-govern [153]. However, after two years and almost three million users, Bluesky's federated protocol only has one server, administered by Bluesky employees, who willingly or not have immense influence over acceptable speech on their platform. By contrast, Mastodon has thousands of federated instances, each with their own moderators and content policies. However, if instance administrators wish to federate with the largest three instances, containing more than half the Mastodon population, they must have a compatible content policy, enforcing an implicit monoculture. These patterns can be identified by measuring the number of communities on a platform,

and the influence that the largest communities have over their peers.

One common approach to measuring community-level centralization is through community size-distribution. If a small oligarchy of Mastodon instances dwarf the population sizes of their peers, then one could presume that the platform is centralized around these instances. Indeed, several prior studies on Mastodon use community size disparity as a starting point, or presuppose that the largest instances are the most significant and focus their study on the largest communities [132, 177, 178, 94]. While the community size distribution is related to centralization, assuming they are the same precludes the possibility that a collection of many smaller instances may be more influential than the few largest, or that the influence of the largest instances may not be directly related to their size.

We reject the assertion that the largest communities must be the most significant, or that their size alone implies centralization, on the grounds that community size does not correlate with the number of cross-community links in observed real-world networks. In fact, our results show multiple platforms where the largest communities are *not* well integrated with the platform as a whole (discussed in results, especially fig. 4.5), allowing a more decentralized network of communities to exist outside of the largest groups. Under this view, the largest communities would be the most significant only when they also act as important information bottlenecks for the entire system.

To illustrate this discrepancy, consider fig. 4.1. In the centralized panel the largest community serves as a central hub, connecting several smaller communities together through shared membership. In the decentralized panel community size is normally distributed, and no community has a pivotal role as a bridge between its peers. Community size-distribution and graph-centric metrics like betweenness-centrality would

agree that the former network is centralized, while the latter is decentralized. However, the third ambiguous panel presents a more complex scenario: the community size distribution is highly unbalanced, but the largest community has almost no impact on the remainder of the network. The largest community has a high betweenness-centrality because of its pivotal role in connecting so many users to the rest of the graph, but it has a long path distance from users in other communities and does not serve as a bridge between communities, and so betweenness-centrality does not match our intuition that the largest community has a small role in the rest of the network.

We propose a definition of centralization meant to capture the alignment between rankings of community size and information bottlenecks. To do so, we combine theoretical ideas from graph theory on bottlenecks and applied concepts from network science about network resilience. Our metric then measures how removing a community would impact users within remaining communities, based on the number of "bridges" between communities. We study a variety of real and simulated networks with this method to examine platform behavior under a range of conditions, and we compare our metric to existing measurements of centralization and network "bottlenecks." Finally, we discuss how this work contributes to broader discussions of centralization online, and how techniques like ours can be extended with richer interaction data.

4.1 Prior Work

Centralization of online platforms is sometimes defined in terms of decision-making power, or who has the authority to make what kinds of decisions about the use of the platform. This definition can be traced to Elinor Ostrom's work on Institutional

Analysis and Development [121], which describes "layers" of decisions, from operational rules (elementary actions any user can perform), to collective rules (the context in which users operate and interact, such as the Twitter feed or the Amazon marketplace), to constitutional rules (the "meta" rules through which the system changes itself). Modern research on platform design often assesses who has decision-making power, and what levers of change are available to different categories of participants [90, 48].

While qualitative studies examine power structures through analyzing governance and rule sets [143, 45], network science infers structure through the observed interactions between humans [47, 117]. We quantify centralization using attributes that fall into three categories: vertex-level attributes, cluster-level attributes, and graph-level attributes. Vertex-level attributes like betweenness centrality [47] or eigenvector centrality [21] measure the prominence of a particular node in terms of how well it is connected to its peers, or how many paths flow through the node. Cluster-level attributes describe groups of vertices, such as the size of the population that contains a particular attribute, or the assortativity describing how likely vertices with a particular attribute are to be connected to one another. Graph-level attributes describe aspects that span the entire network, including diameter, density, and graph conductance [113]. Quantifiability should not be conflated with objectivity; the modeling choice of what entities are included as vertices and what relationships are represented as edges or attributes presupposes what can be considered influential or centralized [24].

Another thread of research tries to join the social theory of centralization and graph theoretical metrics. [45] distinguish between the technical underpinnings of a

network and its social layers, focusing on community-run moderation in infrastructure-centralized (Slack, Discord) and self-hosted (Minecraft) services. Prior Mastodon research also bridges this gap, including both geographic and data-center distribution of instances [132], important for understanding resiliency to disruption or power-outage. This approach aligns with notions of network robustness where centralization can be measured by how a network breaks down under targeted pruning of central nodes [5]. Other studies on Mastodon also integrate its social interaction graph [94], important for understanding the influence of sub-communities and their administrators on discourse. Studies on the social structure of Mastodon primarily focus on individual-centralization, such as a "border-index" of what fraction of a user's neighbors are on a foreign instance [178] and whether some users serve as critical bridges for information flow between instances [95], or community-centralization, such as how clustering coefficients differ between communities (instances) [177]. Our work intends to add to these options, by considering both a community-level centralization metric of how much influence one community has on the broader platform, and a graph-level centralization score of how quickly a network deteriorates as its largest communities are removed, indicating how much it tends towards monopoly or oligopoly.

Recent social media studies highlight the difference between size and importance, demonstrating the need for a better understanding of smaller-yet-influential subgroup dynamics. For example, [17] identifies a single low-follower Twitter user that has a disproportionate influence on national COVID-19 discussion by starting arguments in the replies to the tweets of public officials. Despite not fitting the typical high-follower and high-engagement profile of an "influencer" or "Internet celebrity," this account's behavior and structural role adjacent to prominent accounts leads to outsized impact.

At a regional scale, [50] focuses on sentiment-spreading dynamics between Japanese prefectures, proposing a causal measure of social influence based on correlated sentiment between geographic regions in a forecasting model. Other researchers have focused on cross-platform misinformation campaigns, including [87], showing how bad actors can coordinate across YouTube, Facebook, and Twitter to thwart content moderation. We believe that measuring community-level influence through observed social interlinking will contribute to this conversation on disproportionate influence at multiple scales.

4.2 Methods and Materials

In the following sections we introduce our metric and two data sets: five real world networks that encompass a breadth of configurations, and a set of common synthetic networks.

4.2.1 Measuring Centralization: Disruption Curves

Prior studies on centralization of social networks often focus on graph-level attributes such as detecting components, the size of the giant component, modularity, density, degree distribution [11]. Others may use "bottleneck" metrics like graph conductance [113] to identify bridges and key clusters. These metrics are most appealing in unipartite settings where the structure of the network is not prescribed. However, we focus on bipartite graphs where communities are well defined, such as subreddits, Mastodon instances, or newsgroups. In these contexts, we are not attempting to infer the number or boundaries of communities, but to measure how influential the known

communities are on their neighbors. The size distribution of communities tells us how large a subgroup is, but does not capture the overlap between communities. A graph-wide modularity score describes how well-partitioned the graph is into clusters, and so approximates how insular communities are, but cannot provide more nuance as to whether the largest communities are more integrated than smaller ones, whether small communities are well connected to larger peers but not to each other, or other topological features.

We propose that the influence of a community should be measured in terms of how users outside the community would be impacted by its absence. In other words, a community's influence should be proportional not to its size, but to the number of bridges between it and other communities. Or, in graph theoretic terms, what fraction of edges would be cut by removing a community, not counting users that do not participate outside the community. More succinctly, "what percentage of edges from surviving vertices would be cut by removing a community?"

We measure disruption cumulatively, rather than discretely per-community. This allows us to answer questions like "how influential are the largest three communities on the rest of the platform?" Since "oligarchies" of large and densely interconnected communities may be common, a cumulative metric is more useful than measuring the influence of a single community on the rest of the oligarchy.

Formally, we define a set of communities that are being cut, A, with associated edges $|A|$. Each user has a set of edges to one or more communities. If users *only* have edges to communities in A, then the user is removed along with A. Surviving users with an edge to at least one remaining community are denoted S, with total edges $|S|$, and edges to cut communities in A denoted ∂S. The disruption curve is calcu-

lated as $\partial S/|S|$. This notation was chosen for its similarity to the Cheeger number [113], stressing how our metric measures the alignment of community size and information bottleneck. We additionally outline the algorithm as pseudocode in listing 4.1.

```
disruption = []
for c in communities:
  remaining = 0
  original = 0
  removeCommunity(c)
  for user in users:
    if degree(user) > 0:
      remaining += degree(user)
      original += originalDegree(user)
  disruption += [1-(remaining/original)]
```

Listing 4.1: Pseudocode for disruption algorithm

We recommend caching the size of the smallest community that each user participates in, and pre-sorting users by the order in which they will be removed to avoid computationally expensive references to a graph or adjacency matrix during each removal-step.

Our disruption curve metric is intended for bipartite networks, where communities are clearly distinguishable with ground-truth definition and users can participate in multiple communities. However, some consideration is also given to applying our metric to unipartite settings in the synthetic network section.

We plot disruption similarly to a cumulative disruption function (CDF), where the x-axis represents the number of communities removed, cumulatively ordered by degree, and the y-axis represents the fraction of edges from surviving users that have been cut. In other words, the x-axis is the size of A as a fraction of all communities in the graph, and the y-axis is $\partial S/|S|$, where both the numerator and denominator are dependent on $|A|$.

While disruption curves offer insight into the role of the largest communities on a

platform, some readers may desire a scalar summary statistic to describe how "centralized" a platform is under our metric. For these scenarios we recommend calculating the area under the curve, as shown in figs. 4.3b and 4.4b. We calculate the DAUC using a trapezoidal approximation in logarithmic space. For some synthetic networks it is possible to write a closed-form integral for the disruption curve, but because this is not possible for real-world data, we use a trapezoidal approximation for all real- and synthetic-networks for consistency. We measure the AUC in logarithmic space, because measuring in linear space would heavily weight the influence of the smallest communities that are removed last, and our primary interest is in examining the influence of the largest communities on the broader population. The Disruption AUC (DAUC) does not indicate how much any particular community influences its peers, but summarizes whether a network is prone to disruption if its largest communities are removed.

Platform	**Community Definition**	**Edge Definition**	**Edge weight**
Mastodon	Mastodon Instances	Between each user and every instance on which they follow users	The number of users followed on an instance
Penumbra	A git server	Between a user (identified by email) and each server on which they have contributed to a repository	The number of repositories committed to on each server
BitChute	BitChute channels	Between each user and every channel they have commented on videos from	The number of comments made
Voat	A Voat "subverse"	Between each user and subverses they've participated in	Number of comments made in a subverse
Usenet	A Usenet newsgroup	Between each user and every newsgroup they have posted in	The number of posts made

Table 4.1: Definitions of communities and edges for each platform examined

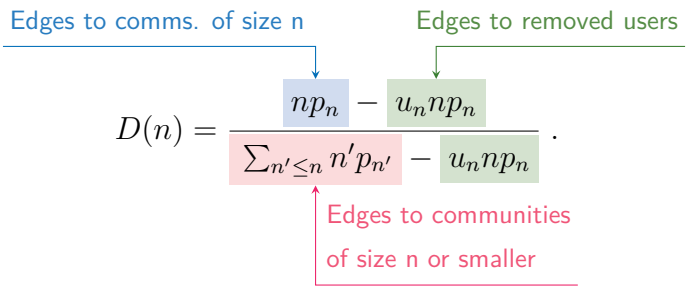
Platform	Comms.	Users	Edges
Mastodon	3,825	479,425	5,649,762
Penumbra	841	41,619	108,038
BitChute	29,686	299,735	11,549,058
Voat	7,515	3,624,486	16,263,309
Usenet	333	2,080,335	58,133,610

Table 4.2: Population size of each network in terms of community count, user count, and relationship edge count, before compressing duplicate edges into weighted edges

4.2.2 Mathematical Analysis of Disruption

We can analyse the expected behavior of disruption curves using random bipartite networks parameterized by their joint-degree distribution. This approach fixes the distribution $\{g_m\}$ of users part of m communities, the distribution $\{p_n\}$ of community size n, and the joint-distribution $P_{n,m}$. Beyond these constraints, we assume the networks to be very large and fully random.

We can calculate the *expected* disruption $D(n)$ involved when removing communities of size $n' < n$. Disruption is given by the number of edges that belong to communities of size n minus the fraction u_n of those that are the sole edge of the corresponding users (since these users are removed in the pruning) divided by the number of edges belonging to communities of size equal or smaller than n minus the $u_n n p_n$ users removed. We write:



(4.1)

The quantity u_n is defined as the probability that a random user of a community of size n has no community smaller than n:

$$u_n = \sum_m \frac{P_{n,m}}{\sum_{m'} P_{n,m'}} \left(\frac{\sum_{n' \geq n} P_{n',m}}{\sum_{n'} P_{n',m}} \right)^{m-1} . \tag{4.2}$$

Fraction of users in communities of size n that have m edges

Fraction of users with m edges in communities larger than size n

In a simple experiment, we create a random Erdős-Rényi-like bipartite network and correlated equivalent networks with the same degree distributions and variable community-user degree matrices $P_{n,m}$. The random network has a simple $P^{\text{rand}}_{n,m} \propto np_n mg_m$ (normalized). We also calculate the maximally assortative $P^{\text{max}}_{n,m}$ by assigning users with highest degrees m_{max} to the largest communities, and maximally disassortative $P^{\text{min}}_{n,m}$ by assigning users with the lowest degree to the largest communities.

Using Eq. (4.1) on networks linearly interpolating between $P^{\text{max}}_{n,m}$, $P^{\text{rand}}_{n,m}$ and $P^{\text{min}}_{n,m}$, we find that positive user-community degree correlations increase disruption and therefore *centralizes* the resulting socio-technical network. Conversely, negative correlations decreases correlations and *decentralizes* the network. We thus know that dispersion curves will be affected by network structure beyond its distribution of community sizes.

4.2.3 Real-World Network Data

We analyze five real-world datasets, each describing online social interactions in bipartite configurations where vertices represent either "users" or "communities." We

utilize a 2021 scrape of the Mastodon follow graph [177]. Mastodon is a Twitter alternative where users are located on one of thousands of "instances," which are Twitter-like servers with their own administrators and content policies. However, Mastodon users can follow users on other instances, exchanging content between the two communities, so long as the servers are "federated" (willing to exchange content). For a second example of a platform with distributed servers, we include the Penumbra of open-source [160], a data set of independent git servers (not GitHub or GitLab), and users that contribute to repositories on each server. We also include an interaction graph from BitChute [162], an alt-tech YouTube alternative, consisting of users and the channels (video uploaders) whose videos they commented on. We utilize a similar scrape of Voat [110], an alt-tech Reddit alternative active until late 2020, consisting of users and the "subverses" (subreddits) they commented in. We additionally include an archive of Polish Usenet groups [151], providing a much older but similarly structured platform for comparison. Details on the vertex and edge definitions for each network are included in table 4.1, and the size of each network is listed in table 4.2.

We selected these platforms because they have clear bipartite user and community representations, data is readily available, and each platform is small enough to obtain a nearly-complete sample. Sub-sampling a larger platform like Reddit may miss lower-population or lower-activity sub-communities, and we are particularly interested in the interactions between smaller communities. The resulting dataset encompasses a variety of approaches to hosting and community governance, providing a spectrum of "centralization."

4.2.4 Ethical Considerations

Any method for community detection, or the measuring of community "importance" (in our case, disruptive potential) in a social context implies risk. A bad actor could use such a metric to identify communities with the most reach in an effort to spread misinformation more efficiently, or to identify the smallest set of content moderators that must be influenced to enforce a desired social policy change across a platform. However, the same methods can be used to preemptively identify risk, allowing a platform like Mastodon to proactively take steps to limit their dependence on their largest server instances, for example by closing user registration on their largest instances, or by de-prioritizing those servers on instance-recommendation websites like instances.social. We believe the benefits of studying the structure of social media outweigh the risks in this regard.

Concerning the datasets used in this paper, we present only aggregate group behavior to provide insight into social welfare. We do not publish any usernames or study individuals' behavior in-depth, and present only the names of some of the largest communities to help contextualize our findings. We believe this presents a minimal risk to the privacy of users included in our five real-world datasets. Since our real-world data is sourced from prior publications we are not re-publishing it with this study, but also do not have the opportunity to further anonymize user data.

4.2.5 Synthetic Network Data

To understand disruption curves and contextualize our real-world results, we examine a variety of well understood synthetic network topologies.

First we construct a bipartite star network, as a default example of a network centralized around a single hub. Bipartite Star networks are analogous to a unipartite star network with duplicate edges. Starting with a unipartite star, we replace each edge from the hub to a leaf with a two-path from the hub community to a new "user" vertex, to the leaf community. Duplicate edges from the unipartite hub to leaves are converted into multiple users that share a community, and serve to break ties when pruning communities for disruption curves. In our example plots, we construct a graph with 150 communities and 3000 users, such that every user has an edge to two communities: the central hub, and one other, assigned uniformly. Removing the hub eliminates 50% of all edges, and removing any subsequent communities incurs no additional disruption, because all impacted users will have a degree of zero and be pruned from the graph (see fig. 4.4a). This graph type is therefore highly centralized but has a decentralized periphery after the removal of the central community, illustrating how different topologies can co-exist in the same network, muddying the definition of "centralization."

We then test disruption on a variety of bipartite networks with power-law degree distributions. We first adapt the Barabási-Albert preferential attachment model to a bipartite setting, initializing a network with 300 empty communities and introducing users that connect to a given community with probability proportional to their size plus one. We also introduce a range of bipartite configuration models: in each, we assign a degree to each community drawn from a power law with a specified γ exponent. For each community, we create edges according to degree, connecting the community to users uniformly randomly without replacement. Therefore, we control for the size of communities, which follow a power-law distribution, but we do not

control for the degree of users, which follow a normal distribution, nor do we control for assortativity. Each of these networks produces a curve that slowly decays towards a diagonal, implying that removing the largest communities has some disproportionate impact, after which removing additional communities has a less pronounced result.

We also adapt the Erdős-Rényi model to a bipartite setting by creating vertices for communities and users, then creating all possible edges with a probability p (in our tests, $p = 0.05$), while preserving the bipartite constraint. These networks produce a disruption curve with a second derivative near zero, indicating that most communities have near-equal influence on the population, and so removing the largest communities does not have a much larger impact than removing subsequent communities.

Lastly, we create a bipartite Watts-Strogatz small-world model. We begin by producing a *unipartite* network with desired neighborhood size ($n = 5$) and edge density ($p = 0.05$) parameters. We apply a clustering algorithm (in our examples we have used weighted community label propagation) to place each user in one community, we create a vertex for each detected community, and we replace all user-user edges with user-community edges. These networks have the most uniform community size distribution of any we tested, and their disruption curves are similar to those of Erdős-Rényi networks, with slightly more variability. By applying community detection, as discussed here and illustrated in fig. 4.2, it is possible to measure disruption in unipartite networks as well as bipartite.

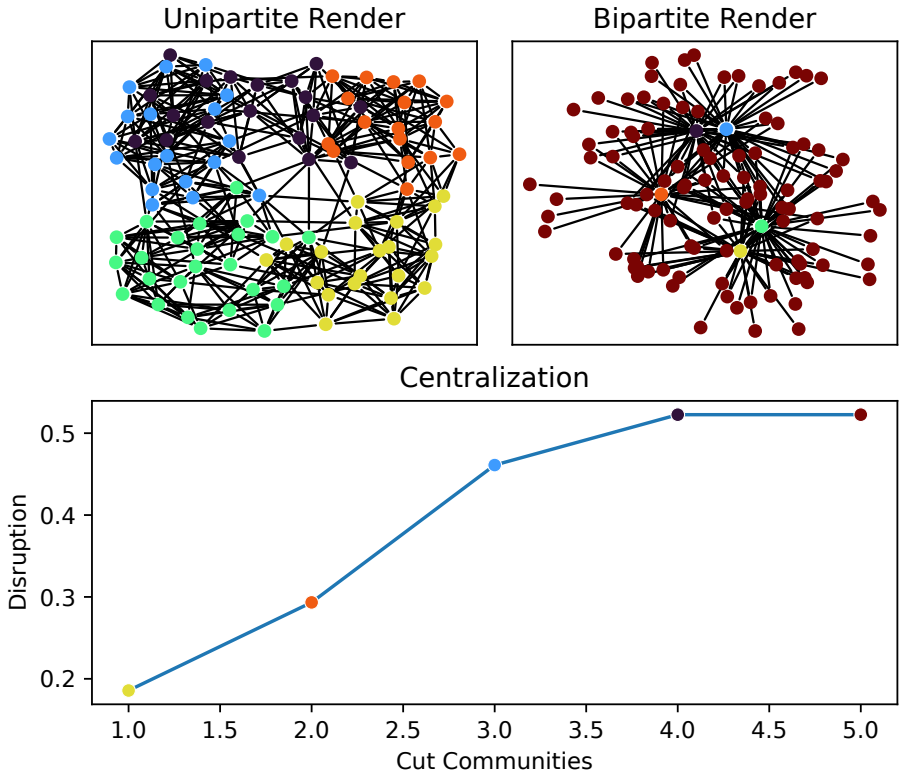
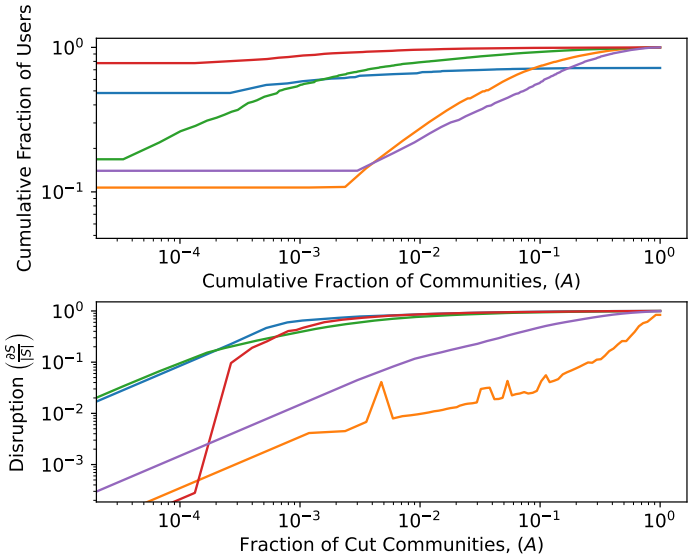


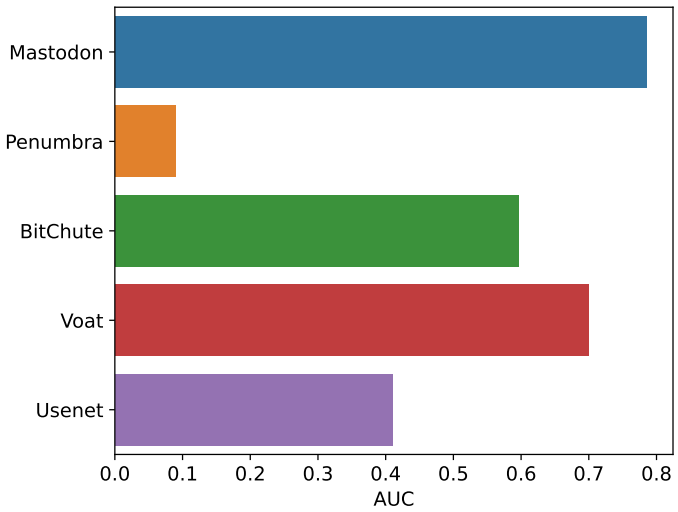
Figure 4.2: Example of applying our disruption metric to unipartite graphs by detecting communities on a unipartite small-world network (top-left), converting labeled communities into a bipartite representation (top-right), and running our influence metric on the bipartite graph (bottom)

4.3 Results

We plot the cumulative population size, disruption curve, and disruption AUC for real-world networks in fig. 4.3, and plot the same results for synthetic network data in fig. 4.4. We first focus on discrepancies between the size distribution and disruption curves for real networks, then return attention to synthetic network data when we examine the role of assortativity.

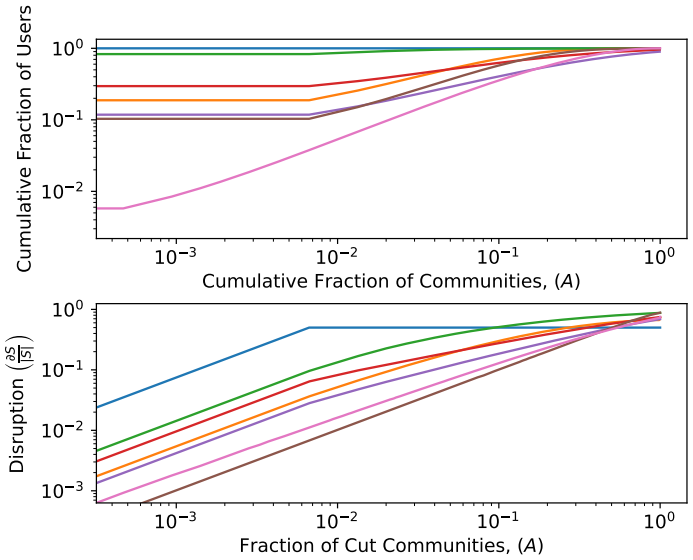


(a) Community size distribution and disruption curves

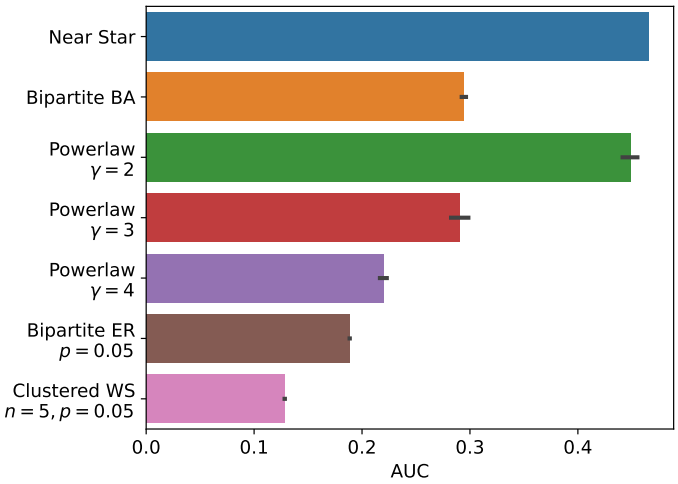


(b) Area under the disruption curve (DAUC)

Figure 4.3: Summary measures of centralization. (a) The population distribution of communities (top) does not correlate with our measure of community disruption (bottom). (b) The area under the disruption curve (DAUC) provides a summary statistic of the disruption curve that reinforces how network structure combined with community size provide greater insight into centralization. Panel (a) consists of cumulative distribution plots of population and disruption, where the top subplot is a CDF of the platform population as smaller communities are included, and the bottom subplot shows how networks are damaged as more of the largest communities are removed. Each line represents a different network, using the color key from panel b.



(a) Community size distribution and disruption curves



(b) Area under the disruption curve (DAUC)

Figure 4.4: In simulated networks with a variety of degree distributions, the disruption curves for each network much more closely match the population distribution (fig. 4.4a), suggesting that non-degree network attributes such as assortativity play a crucial role in determining centralization. As in fig. 4.3, the left figure represents cumulative population and disruption as more communities are considered. Each line represents a network sharing the color-key in the right figure. Simulated networks were generated 100 times, and the mean and a 95% confidence interval are shown in both figures.

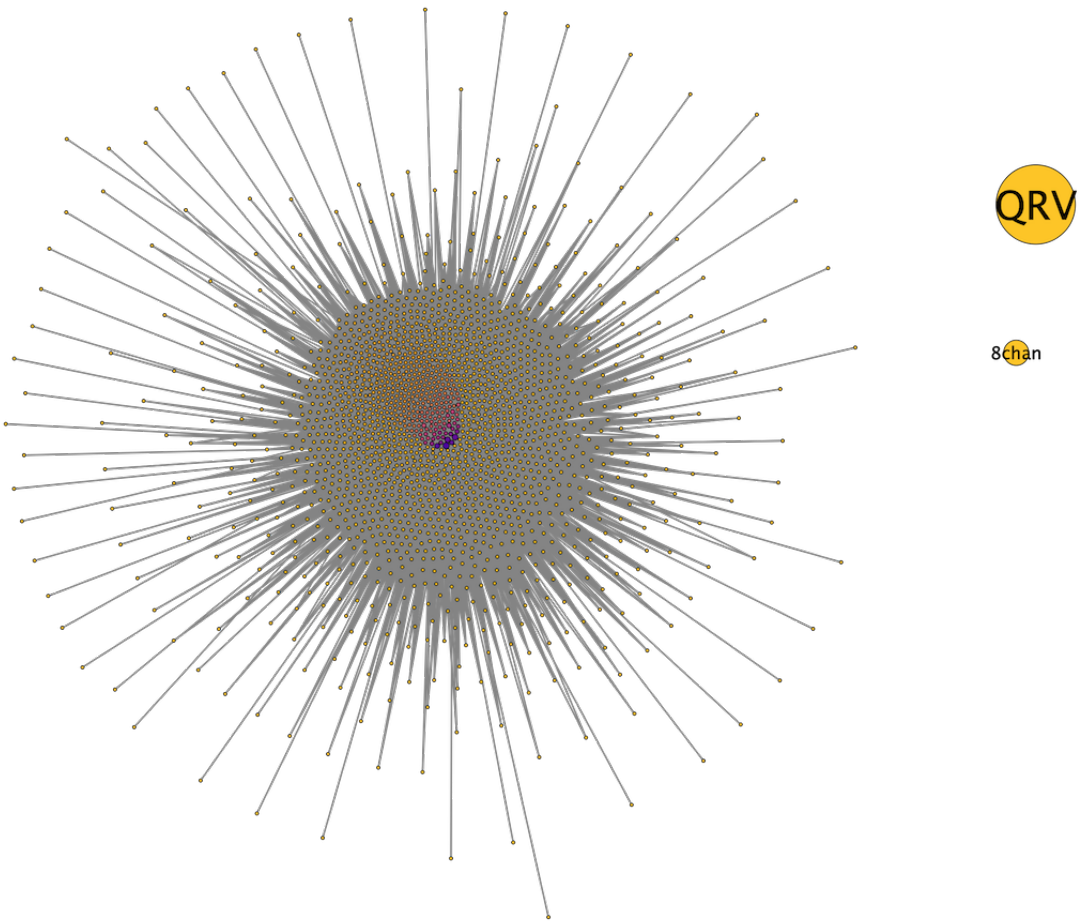


Figure 4.5: The two largest Voat communities ('QRV' and '8chan') are dramatically larger than their peers, but have almost no overlap in population, making community size a poor proxy for platform-wide influence or centralization. In this network visualization, nodes represent Voat "subverses," and edges represent at least thirty shared users active in two communities. Node size correlates with user count, and color correlates with strength; i.e. the level of overlap with neighboring communities. The purple communities at the center are default subverses all new users are subscribed to ("news," "whatever," etc), surrounding pink and orange communities are popular with lots of user overlap. The largest two communities, "QRV" and "8chan," have almost no user overlap with other communities and are rendered to the right.

4.3.1 Comparison to Size Distribution

Upon comparing the size distribution and disruption curve in fig. 4.3a, it is apparent that the community size distribution is insufficient to describe the structure of a network. Voat has the most skewed population distribution: almost all users participate in the largest community, yet the network does not experience significant disruption until the largest *three* communities are removed. Mastodon and BitChute have the next most skewed size distributions, but there is a large distance between the proportional sizes of their largest communities, and almost identical disruption curves as those communities are removed. By population distribution, the Penumbra appears to be more skewed towards its largest git servers than Usenet is towards its largest newsgroups. This is not mirrored in disruption curves, where Usenet has a consistently higher disruption than the Penumbra.

To explain these discrepancies, we examine each network in greater detail. Voat was a Reddit-like platform where users commented and posted in one or more "subverses." While users chose to subscribe from among 7515 public subverses, new accounts were automatically subscribed to a set of 27 subverses by default. This "default subscription" has no parallel on other platforms we examined. Since these default subverses have an automatic population, they are more likely to receive engagement than subverses that must be discovered according to a user's area of interest, and we may expect them to be densely connected with most users on the platform. However, the largest two subverses on Voat by number of unique users were *not* default subverses; `v/QRV` was a QAnon conspiracy group, and `v/8chan` was a right-wing news and discussion forum whose name references the white supremacist imageboard 8chan

(now “8kun”). Both subverses were highly insular, with little population overlap with the rest of the platform, as illustrated in fig. 4.5. Therefore, it is only when we remove the *third-largest* subverse, `v/news`, that we see a large impact on remaining users on the site.

The Penumbra of open-source represents software development on git servers outside of GitHub and the primary GitLab instance. Each community represents a git server with one or more public repositories, and edges indicate that a user (identified by email address) contributed to a repository on a server. Servers are often created per-organization; for example, the Debian Linux distribution hosts their own GitLab server at `salsa.debian.org`. Users often contribute to multiple repositories on a single server, but connections *between* servers are extremely sparse. This sparsity is responsible for the “spikes” in the Penumbra’s disruption curve; removing a git server may sever an edge to some users, and removing a second, related server may prune all remaining edges to those same users. When the cross-server collaborative users are removed, the impact on the remaining less-collaborative community decreases. In all other networks enough users have a sufficiently high cross-community degree that disruption only increases as communities are removed.

4.3.2 Comparison to Giant Component Size

Rather than examining the cumulative community size distribution, one could instead examine the size of the giant component of each network. The giant component will shrink as communities are cumulatively removed, providing another means of examining the influence of large communities.

We illustrate this cumulative shrinking in fig. 4.6. Most curves are smooth until the

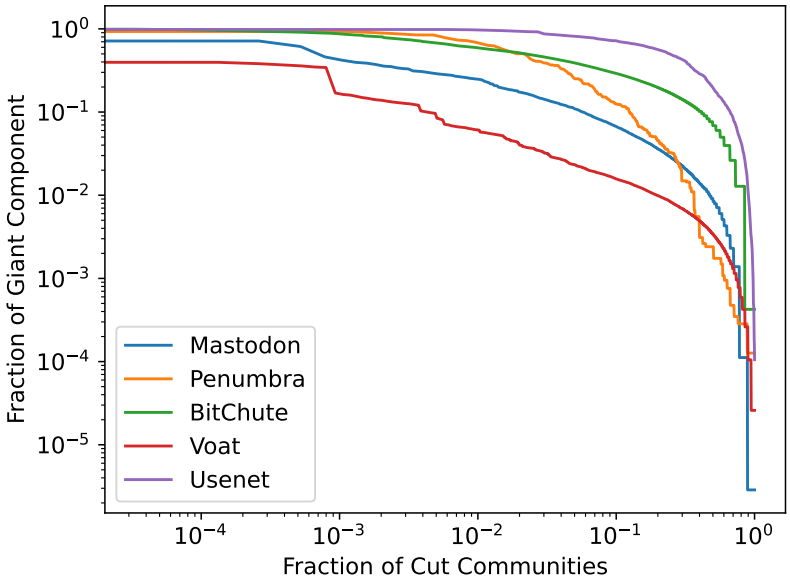


Figure 4.6: The giant component shrinks as communities are pruned from largest to smallest. Line slope indicates both the size of a community and whether it was part of the giant component before pruning. However, boolean inclusion does not account for how well-integrated the community was among its peers. The y-axis is normalized as a fraction of the un-pruned giant component size, such that "0.5" indicates the giant component is half the size of the original.

tail of the distribution, with two notable exceptions: Voat's giant component changes once the largest insular communities are removed (see fig. 4.5), and the Penumbra's curve is much "spikier" as a result of its highly sparse structure.

Measuring the change in giant component size captures some of the same features as our disruption metric. In particular, removing large insular communities may not change the giant component size if the community is completely isolated from the giant component. However, the impact of a community is boolean: if it touches the giant component, then removing the community will shrink the giant component by the size of that community. There is no distinction between a minimally integrated and tightly integrated community. Measuring the impact of a community in terms of fraction of edges severed, rather than component vertex size, offers finer insight into the interplay between size distribution and network structure.

4.3.3 Comparison to Network Bottlenecking

The Cheeger number [113] is a single-valued metric representing how large of a "bottleneck" inhibits conductance across a graph. It is a minimization problem that seeks to divide vertices into two large clusters with a small number of links between them, which is similar to maximizing two-partition modularity. It is typically written as:

$$\min\left\{\frac{|\partial A|}{|A|} : A \subseteq V(G)\,,\ 0 < |A| \leq \frac{1}{2}|V(G)|\right\} \tag{4.3}$$

Edges crossing the boundary of A

A is a subset of vertices of G

All edges in+across A

A contains at most half of all vertices

Graph conductance is a global search that measures how similar a graph is to a "barbell," where a small score indicates large communities with few edges across the bottleneck. We are also interested in the size of the bottleneck created between large communities and the rest of a platform, where a large bottleneck and low number of edges among "surviving" users indicates high disruption. However, while the Cheeger number is a community-search algorithm, the communities in our bipartite social-network setting are predefined, and we are interested specifically in the size of the bottleneck for surviving users when the largest communities are removed. Our partition selection is bipartite-aware, such that A includes all the largest communities we are pruning, and all users that only have edges to those communities. Additionally, while the Cheeger number returns a single value for the most "barbell-like" partitioning the graph can achieve, we are interested in the cumulative effect of pruning more and more communities as a means of identifying oligarchic patterns in a network.

Unfortunately, evaluating the graph conductance of all possible subsets of vertices is an NP-hard problem [83] such that it is impractical to directly measure the Cheeger constant on most large graphs. The Cheeger inequality offers upper and lower bounds on the Cheeger number based on the second eigenvalue of the normalized Laplacian of the adjacency matrix, but in our tests these bounds were too wide to offer insightful comparison.

4.3.4 Assortativity and Centralization

High degree disparity is not enough to create a network as centralized as Mastodon. When we control for degree distribution using a variety of "centralized" models including star networks and powerlaw distributions we cannot achieve more than 50%

disruption (fig. 4.4b). To achieve higher disruption you must have duplicate edges, representing for example a Mastodon user following many accounts on the same server. Therefore, we expect that degree assortativity (or degree-degree correlation) plays a significant role in the differences between observed community disruption (fig. 4.3a) and network behavior under controlled degree distributions (fig. 4.4a). In a purely random setting, users are likely to have edges to multiple large communities, because most edge stubs in a configuration model come from high-degree communities. In real social settings, the content of communities may inhibit assortativity, as in Voat, where the largest two communities are highly insular (see fig. 4.5), creating a large disparity between the community size distribution and disruption metric.

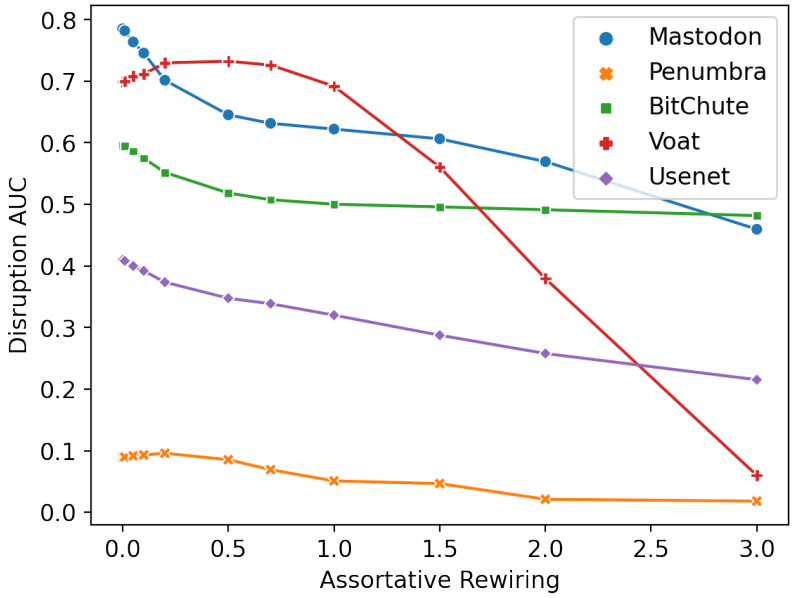


Figure 4.7: Increasing user-community degree assortativity through edge-rewiring increases the influence of the largest communities in highly insular (Voat) or sparse settings (Penumbra), but decreases disruption in all networks as increased rewirings eliminate cross-community edges and yield insular and sparse networks. Y-axis represents disruption AUC (see fig. 4.3b), so that the slope shows change in disruption AUC as networks are rewired to increase user-community degree assortativity.

To explore this hypothesis, we randomly rewired each social network to increase assortativity. We select pairs of edges uniformly without replacement, and swap the communities of the edges if it would increase user-community degree assortativity. We continue this process until we have rewired a desired percentage of edges; if we exhaust the edge supply before finding sufficient valid swaps, we re-shuffle the edge list and continue drawing. For each rewired network we calculate its disruption and the area under the disruption curve, as in fig. 4.3b, and plot the change in AUC during rewiring in fig. 4.7.

This experiment is useful in distinguishing the idea of network centralization from classic ideas of monopoly. These are two different, but related, problems that are easy to confuse when focusing solely on summary statistics like community size distributions. When a network consists of disconnected communities, it is decentralized under the disruption metric regardless of the size distribution of these communities. This conclusion follows from our definition of centralization since removing a community in a sparse (or disconnected) network, has little (or no) impact on other communities. This rewiring experiment highlights this logic: As networks get rewired to increase correlations, we increase the likelihood of having all the activity of a user focused on a single community and therefore progressively disconnect the community and decentralize the network. The only exception is Voat, whose initial state contains large disconnected communities that can get coupled to the rest of the network by rewiring, before being re-disconnected as we rewire more and more. Small correlations in large networks can therefore increase centralization, since large communities can broker more bridges when they contain well-connected users; while strong correlations in smaller networks can decrease centralization by focusing user activity on

single communities.

There are multiple interpretations of degree assortativity in a bipartite setting. The linear correlation between user degrees and community degrees measures whether high-degree users are likely to be connected to high-degree communities. In our network definitions edges represent activity, like follow relationships or participation in conversations, so this measures whether active users are likely to be connected to communities with lots of activity. A second metric of interest is whether large communities are likely to be connected to other large communities, or the assortativity of a unipartite-projected community-community graph. This can be broken into two sub-cases: assortativity of community size (do communities with many users share users with other high-population communities), and assortativity of degree (do communities with lots of activity share users with other high-activity communities). These three notions of assortativity may correlate if high community population correlates with high activity, but this is not guaranteed, so the three metrics should be measured separately.

While rewiring to promote user-community degree assortativity we also plotted the changes in community-community degree assortativity, shown in fig. 4.8. Strikingly, the community assortativity *decreases* as we rewire to promote user assortativity. This is because as we rewire edges to focus user connections on the largest communities we implicitly decrease the number of edges between communities. This also matches the changes in disruption in fig. 4.7: increasing assortativity may reconnect large and insular communities with the rest of the network, briefly increasing their influence, but continued assortativity rewiring also cuts bridges to and between smaller communities, yielding a sparse network that is far less centralized.

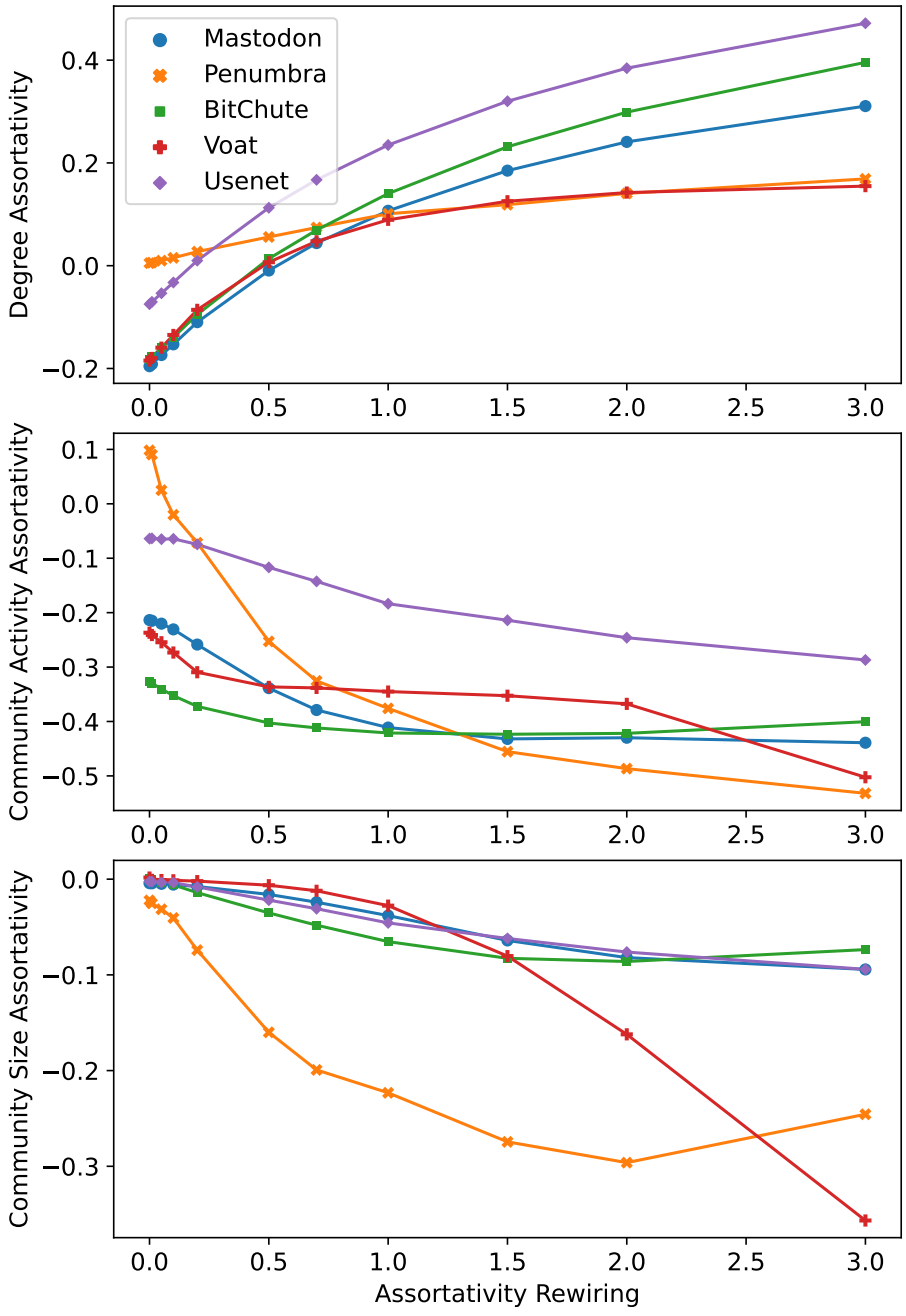


Figure 4.8: Rewiring to increase user-community degree assortativity (top) decreases the projected community-community degree assortativity (middle) and community-community population assortativity (bottom).

To further explore the relationship between these types of assortativity, we also rewired networks in the reverse direction: for randomly selected pairs of edges, we rewired those edges to *decrease* user to community activity assortativity. We have plotted the change in disruption curves (fig. 4.9). In most networks, decreasing activity assortativity lowers centralization, although the effect diminishes as the network topology more closely approximates a random network. The one exception is the Penumbra; this network has such sparse inter-community connections that any perturbation of edges increases the cross-community links, community-activity assortativity, and community-size assortativity, and therefore *increases* centralization.

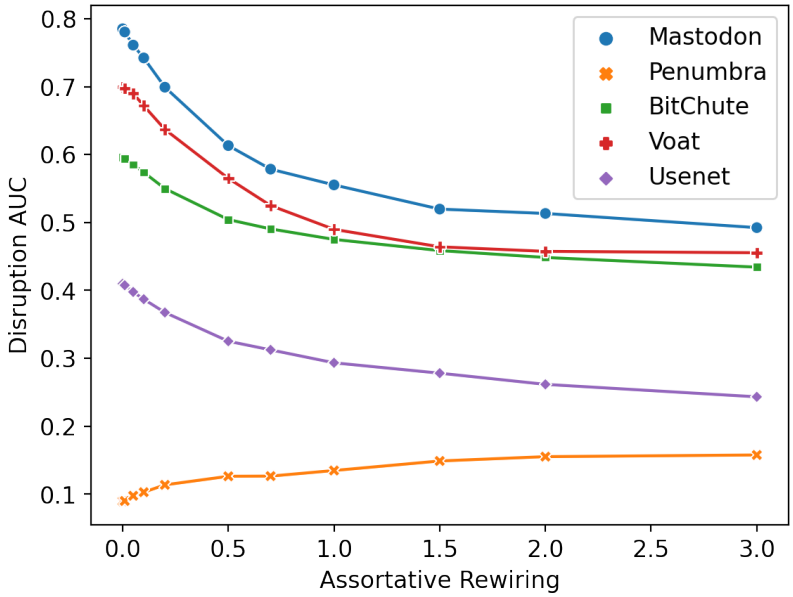


Figure 4.9: Rewiring networks to decrease user-community degree assortativity also typically decreases disruption when large communities are removed. However, for very sparse networks like the Penumbra, any perturbation, including rewiring to decrease assortativity, increases community inter-connection and so increases the influence of large communities.

4.4 Conclusion and Future Work

We have added to the wealth of centralization metrics by proposing a mesoscale measurement that indicates how much influence one sub-community has over a broader network, by accounting for how many edges to remaining users would be severed if a community were removed. This metric allows us to differentiate between networks with a substantial community size-imbalance, and networks where the largest communities play a core structural role in their smaller peers. We extend our metric to create a graph-level measurement that indicates how "oligopic" a network is, or how well-integrated its largest communities are with the population at large.

We assert that a more nuanced measurement of community influence, accounting for both size distribution and structural role, has utility for content moderation and administrative transparency. Identifying communities on a platform with disproportionate influence can help moderators and administrators limit that influence through changes in content recommendation and integration. Conversely, identifying large communities with lower influence than expected can aid in detecting a large influx of external users, as in the QAnon communities on Voat. Furthermore, third party analysis of community influence can reveal when platform administrators overstate claims of decentralization and community self-governance, understating their own control over and responsibility for a platform.

We have utilized our disruption metric to examine a range of real-world social networks, comparing their network topology, distribution of community sizes, and the influence of those communities. We find that some platforms, like Voat, are much less centralized than their skewed community-size distribution would suggest, while

others, like Usenet and the Penumbra of Open-Source, have similar size distributions and widely divergent disruption curves. Mastodon, while vocally supportive of decentralization, has a disruption curve mostly characterized by the skewed population distribution of its sub-communities and is in fact more centralized than any other real or synthetic network considered in this study.

Using simulated networks with a range of degree distributions, and rewiring techniques to adjust assortativity, we have begun to explore the interplay between community size, structure, and community-level centralization. However, we limited ourselves to traditional network generative models like Erdős-Rényi and power-law configuration model networks. Future research could directly simulate networks with chimeric centralization which combine decentralized and centralized components to more realistically represent the diversity observed in social networks.

Our network representations are oversimplified in that we assume that each edge on a network represents a path of information flow. However, one user following another represents *potential* information flow; a bridge between two communities is only realized if the following user is online and chooses to propagate information from the edge to their own followers and instance.

More thorough research should examine how many potential bridges are utilized by, for example, monitoring the number of "boosts" (Mastodon's equivalent to "retweets") across instance boundaries on Mastodon. Observed information spread, and examining the reception of cross-pollinated ideas in non-originating communities, would provide much greater insight into how multi-community platforms function in practice.

Chapter 5

Distinguishing In-Groups and Onlookers by Language Use

Foreword

Online social platforms do not exist in isolation, but are part of a shared ecosystem, frequently linking to one another, sharing screenshots from one another, and moving users and ideas between one another. Therefore, a rigorous study of social and technical platform design should not consider a platform alone, but consider how it interacts with adjacent platforms that offer different affordances.

While I have studied behavior on permissive "alt-tech" platforms [158, 161, 31], studying the migration between mainstream and alternative platforms presents a number of challenges. Primarily, we need to identify that two groups on different platforms are a single shared community. Were all account information public, we may be able to correlate email addresses used to register accounts, or IP addresses from which accounts were accessed, to associate two accounts on different platforms.

In the absence of such explicit identifiers, some studies check for matching or near-matching usernames between platforms; but two similar usernames do not guarantee the same person holds both accounts, and people frequently use different usernames on different sites, making this approach prone to false positives and negatives.

I have proposed identifying communities using a *communal linguistic fingerprint.* Rather than demonstrating that two users are the same person, we could attempt to demonstrate that two groups of users speak about the same subjects using the same in-group vocabulary in the same contextual way. This approach bypasses the more challenging task of identifying corresponding users across two platforms, and could identify shared communities even if there has been some turnover in population. This kind of group-identification is also less privacy-invasive, in that it does not attempt to track or deanonymize individual people, but only show that two large groups of people talk similarly.

Such a linguistic fingerprint would need to overcome several problems. First, it would need to sufficiently capture word-usage and context so as to distinguish between members of a community and members discussing a community. For example, the QAnon conspiracy movement has a range of distinguishing in-group vocabulary and prominent keywords, but people talking about the QAnon movement may use much of the same vocabulary while explaining or debunking the conspiracy. It is this challenge of distinguishing between group members and onlookers of a group that I address in this chapter.

Second, the accuracy of a linguistic fingerprint will decay with time, because in-group vocabulary often centers around current events. For example, QAnon was once focused on "pizzagate," a conspiracy theory that flourished in 2016 and re-emerged

in 2020, which claimed that prominent members of the Democratic Party were part of a pedophilia ring frequently asserted to be in the basement of the Comet Ping Pong pizzeria. In 2020, terms associated with Pizzagate were strong indicators of QAnon affiliation, but those same terms have since diminished in prominence. It may be possible to minimize this vocabulary-shift-driven classifier decay by taking a long time-scale sample of a community and identifying only the distinguishing terms that remain most consistent throughout the sample to create a stable fingerprint. This extension is outside the scope of my research so far, but is critical for observing inter-platform migration, because a move after deplatforming implies a time delay between the text sample of the community available from the old platform, and the text of the community available after they have reestablished elsewhere.

Abstract

Inferring group membership of social media users is of high interest in many domains. Group membership is typically inferred via network interactions with other members, or by the usage of in-group language. However, network information is incomplete when users or groups move between platforms, and in-group keywords lose significance as public discussion *about* a group increases. Similarly, using keywords to filter content and users can fail to distinguish between the various groups that discuss a topic—perhaps confounding research on public opinion and narrative trends. We present a classifier intended to distinguish members of groups from users discussing a group based on contextual usage of keywords. We demonstrate the classifier on a sample of community pairs from Reddit and focus on results related to the COVID-19 pandemic.

5.1 Introduction

Online communities today have unprecedented power to impact the course of disease spread [129, 12], sway elections [23, 127], and manipulate global markets [8]. However, studies of online communities are often limited to single platforms due, in part, to the fact that the overlap in users across platforms is never explicitly known or because user networks and user behavior may differ across platforms [65, 159, 61]. Nevertheless, there are some exceptions (*inter alia* [169, 3, 76]) and account mapping is an area of active research (*inter alia* [28]).

A powerful alternative to account mapping is to track language rather than users, which only requires data on the content of the platform and not necessarily their user base. There remain important caveats to this approach, however: 1) shifts in language can be hard to differentiate from shifts in user demographics and 2) language *about* a group of interest can look very similar to the language *of* the group itself. This is especially true if in-group vocabulary is used by outsiders when discussing the group, or if the in-group's vocabulary percolates into the general lexicon. An example of such language spread involves the word "incel," which was popularized in a specific online community before becoming more widely known.

Here, we address the second problem of distinguishing in-group members from onlookers engaged in discussion about the in-group, based on language alone. We introduce a group-classifier, which labels users as being in a group or discussing a group. We train our classifier on Reddit, an online forum broken into explicit sub-

communities (i.e., "subreddits"). We identify pairs of subreddits, where one subreddit focuses on a particular topic (e.g., COVID conspiracies), and a second subreddit of "onlookers" discusses the first community or topic. Consistent user participation in a subreddit implies group membership, providing training labels; we filter outlier users who participate in or "troll" their chosen subreddit's counterpart. Our classifier attempts to distinguish users from each community based on their usage of topic words.

Our contributions in this piece are focused on two main points:

1. We propose a framing for in-group and onlooker discussion communities and discuss the value of differentiating between them in downstream analyses. This point is especially important for future work on cross-platform community activity.

2. We collect a novel data set of in-group and onlooker subreddit pairs and present a baseline classification pipeline to demonstrate the feasibility of separating groups of users accounts based on the content of their posts. We go on to present preliminary results on how this automatic labelling of user accounts may affect downstream analyses relative to the ground truth data.

The rest of this manuscript is organized as follows: in section 5.2 we provide an overview of prior work, mainly in the complimentary spaces of stance detection and counter speech. In section 5.3 we outline our methods, including the collection of a novel dataset of subreddit pairs. In section 5.4 we present the results from our in-group and onlooker classifier along with the impact of automatic labelling on resulting language distributions. We discuss the implications of our work in section 5.5

and concluding remarks in section 5.6. Finally, in section 5.7 we suggest areas for future work which could build upon our in-group and onlooker framing, improve our classification pipeline, and address broader research questions.

5.2 Previous work

We classify authors as being "in a group", or "discussing a group", not necessarily in an adversarial way. This closely resembles stance detection [92, 7]. Research involving stance detection may be divided into two main categories [7]:

1. Predicting the likelihood of a rumor being true (i.e., rumor detection) by examining whether the stance of posts is supporting, refuting, commenting on, or questioning the rumor [182, 183, 66].
2. Assessing whether the stance of a post is "pro", "against", or "neither" with respect to any given subject [9, 13, 82, 1, 7].

In some cases, manually labelled datasets are used to evaluate the quality of stance detection pipelines [81] or train stance classifiers using supervised learning [114].

Similar to the latter category of stance detection, topic-dependent argument classification in argument mining also parallels our classification scheme, as it may work to evaluate whether a sentence argues for a topic, argues against a topic, or is not an argument [109, 135, 98].

"Perspective identification" works to assess an author's point of view, e.g., classifying individuals as "democrats" or "republicans" based the content of their post [103, 167, 146, 19]. Our work also relates to the automated identification of "counter-speech", in which hateful or uncivil speech is countered in order to establish more

civil discourse [168, 69].

Our work is similar to the form of stance detection that evaluates "pro", "anti", or "neither" attitudes, but the problems of stance detection tend to assume that any discussion about a group are adversarial. However, the problem of distinguishing the language *about* a group from language *of* the group is much more general, as people discussing an emerging subculture do not necessarily oppose it. For example, onlookers may talk about non-political groups formed around new music scenes, small social movements or communities surrounding specific activities without holding opposing views to these groups. Political or not, identifying these onlookers can be of critical importance when studying a specific subculture.

5.3 Methods

5.3.1 Data Selection

Reddit partitions content into "subreddits": forums dedicated to a particular topic, with individual community guidelines and moderation policies. We identified seven (7) pairs of subreddits where one subreddit was focused on a highly-specific topic and another subreddit was dedicated to discussion about the first community. We selected clearly distinguishable communities that formed pairs of in-group and onlooking group subreddits. For example, NoNewNormal is a COVID-conspiracy and anti-vaccination group, while CovIdiots is dedicated to discussing anti-vaccination and COVID conspiracy theories (see Fig. 5.1 for an overview of 2-gram distributions for these subreddits). We selected this pair as our main case study because of the

timeliness of the COVID-19 topic and the volume of conversation in each community. Partially owing to the contentious nature of the communities we were interested in, many of the subreddits we examined had previously been banned. Since data from banned subreddits remains available [15], this did not inhibit our study or reproducibility.

Relationships between the primary community and the onlooking community were typically antagonistic. However, this does not mean that the results from standard sentiment analysis would have been able to correctly classify utterances from each group. For example, the NoNewNormal community may express negative opinions about vaccines or masking mandates, while CovIdiots may express positive sentiment about both topics, but negative sentiment about the opinions held by members of NoNewNormal.

For some of our subreddit pairs, the onlooker subreddit was created specifically to discuss the in-group subreddit. For example, TheBluePill was created in response to TheRedPill. For other pairs, both subreddits discussed the same topic from different viewpoints but were not directly connected. For example, ProtectAndServe is a subreddit populated by current and former law enforcement officers, while Bad_Cop_No_Donut is a subreddit dedicated to the criticism of law enforcement, but it is not specifically a criticism of ProtectAndServe itself. Including both types of subreddit pairs allowed us to measure the effectiveness of our classifier on communities with varying degrees of similarity.

5.3.2 Subreddits Chosen

The following are qualitative descriptions of each subreddit pair we examined. The size of each subreddit corpus, in terms of users and comments, as well as the mean comment score on each subreddit, can be found in the appendix (table 5.4).

`r/NoNewNormal` and `r/CovIdiots`

NoNewNormal self-described as discussing "concerns regarding changes in society related to the coronavirus (COVID-19) pandemic, described by some as a 'new normal', and opposition to [those societal changes]." Most posts focused on perceived government overreach and fear-mongering. Reddit banned the subreddit on September 1st, 2021.

CovIdiots is dedicated to "social shaming" of covid conspiracy theorists, "anti-maskers," and "anti-vaxxers."

`r/TheRedPill` and `r/TheBluePill`

TheRedPill is a "male dating strategy" subreddit, commonly associated with extreme misogyny and a broader collection of "Manosphere" online communities including incels, men's rights activists, and pick up artists.

TheBluePill is a satirical subreddit targeting content from TheRedPill.

`r/BigMouth` and `r/BanBigMouth`

BigMouth is an online fan community that discusses the Netflix television series, "Big Mouth." The show often features coming of age topics, including puberty and teen sexuality.

BanBigMouth was a community focused on associating the TV show with pedophilia and child grooming, and petitioning for the show to be discontinued and

removed. Reddit banned the subreddit in June, 2021 for promoting hate.

r/SuperStraight and r/SuperStraightPhobic

SuperStraight was an anti-trans subreddit that defined "Super Straight" as heterosexual individuals who were not attracted to trans people. Reddit banned the subreddit for promoting hate towards marginalized groups in March, 2021.

SuperStraightPhobic was an antagonistic subreddit critiquing the users, posts, and intentions of the SuperStraight subreddit. It was banned shortly after SuperStraight.

r/ProtectAndServe and r/Bad_Cop_No_Donut

ProtectAndServe is self-described as "a place where the law enforcement professionals of Reddit can communicate with each other and the general public." Users who submit documents proving their active law enforcement status have identifying labels next to their usernames.

Bad_Cop_No_Donut is a subreddit for documenting law enforcement abuse of power and misconduct. Most posts are links to news articles, while comments discuss article content and general police behavior.

r/LatterDaySaints and r/ExMormon

LatterDaySaints is an unofficial subreddit for members of the Church of Latter-Day Saints. While non-members of the church are permitted to ask questions and engage in conversation, criticizing church doctrine, policy, or leadership is forbidden, and the subreddit is heavily moderated.

ExMormon is a subreddit for former members of the Mormon church to discuss their experiences. Posts are typically highly critical of the church.

r/vegan and r/antivegan

Vegan is a broad vegan community, with topics ranging from cooking tips, to

animal cruelty, environmental impacts of meat consumption, and social challenges with veganism.

Antivegan is ideologically opposed to veganism. Much of the subreddit's content is satirical, or critical discussion about the actions of perceived vegan activists.

5.3.3 Data Collection

For each pair of subreddits, we first chose an "ending date" for data collection: If either subreddit was banned prior to the start of our study, we used the earliest ban-date as our ending date. Otherwise, we used the date of our data download. We then downloaded all comments made in the subreddit for one year prior to the ending date, using pushshift.io, an archive of all public Reddit posts and comments which is frequently used by researchers [15]. We then filtered out comments made by bot users, using a bot list provided by [159].

We anecdotally observed users from some of our selected subreddits "raiding" other selected subreddits. For example, users from subreddits opposed to the `r/NoNewNormal` COVID-conspiracy group sometimes harassed users in `r/NoNewNormal`, and vice-versa. We did not want these harassment-comments to bias our text-analysis, so we filtered out all users who had an average comment-score less than unity for their comments in the subreddit. In other words, we only kept comments from users that the community did not strongly disagree with. This did not filter out coordinated attacks, where many members of one community raided another, upvoted their raiding comments, and downvoted the in-community comments. However, this type of attack (often referred to as "brigading") is a bannable offense on Reddit, and we did not observe it in our dataset.

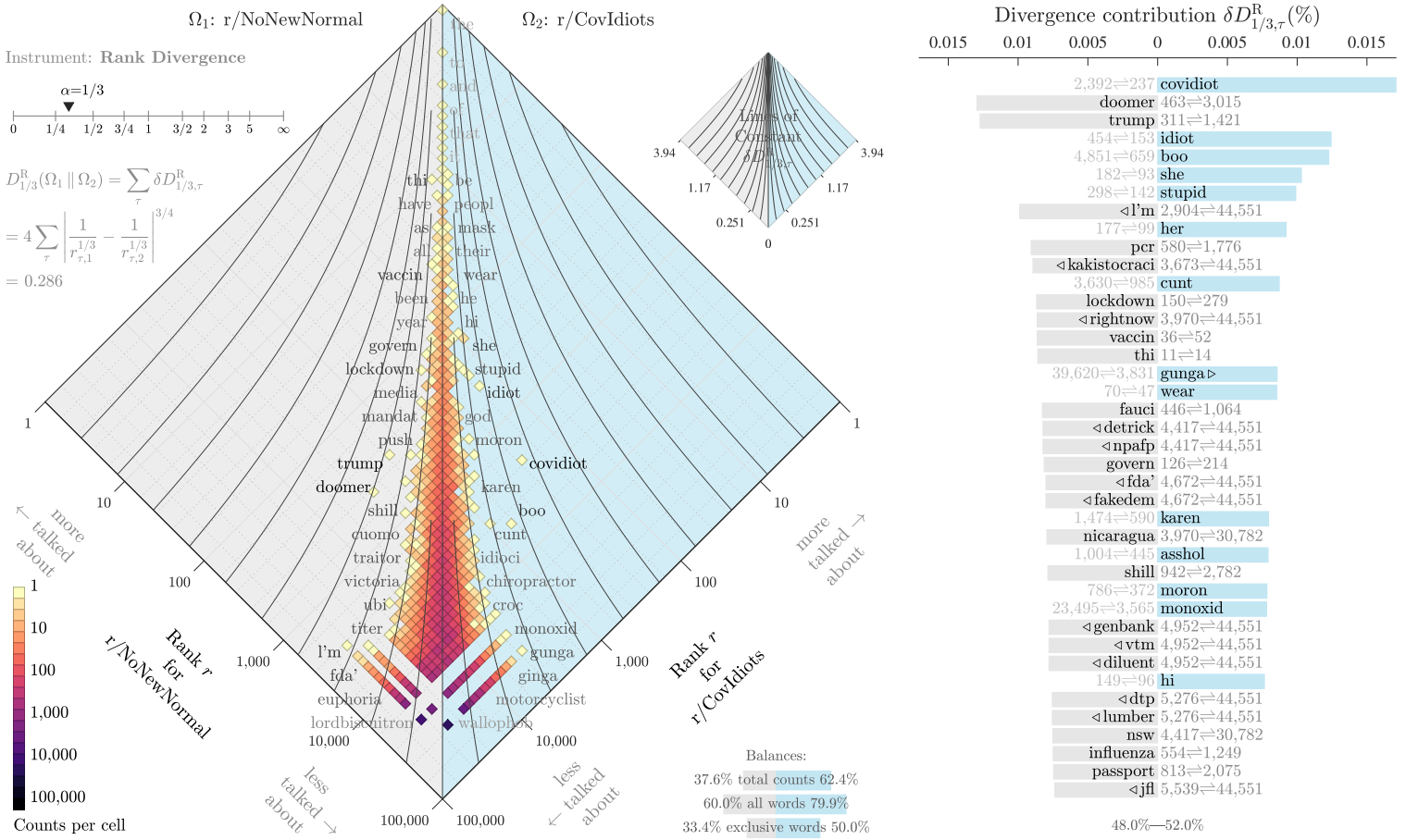


Figure 5.1: An allotaxonograph [41] showing the 1-gram rank distributions of `r/NoNewNormal` and `r/CovIdiots` along with rank-turbulence divergence results. The central diamond shaped plot shows a rank-rank histogram for 1-grams appearing in each subreddit. The horizontal bar chart on the right shows the individual contribution of each 1-gram to the overall rank-turbulence divergence value ($D^{R}_{1/3}$). The 3 bars under "Balances" represent the total volume of 1-gram occurring in each subreddit, the percentage of all unique words we saw in each subreddit, and the percentage of words that we saw in a subreddit that were unique to that subreddit.

5.3.4 Determining In-Group Vocabulary

To compare the n-gram distributions of pairs of subreddits we used rank-turbulence divergence (RTD) [41]. We used RTD to both summarize overall divergence and highlight specific n-grams that contributed most to this divergence value. We found RTD to be an effective choice when making more nuanced comparisons between the disjoint distributions of subreddit pairs. It avoids construction of the mixed-distribution found in other divergence measures—such as Jensen-Shannon divergence (JSD)—which may be less effective at highlighting salient terms with the subreddit-scale distributions. Rank-Turbulence Divergence is described in more detail in section 1.1.4 and section 1.1.4.

We used a divergence-of-divergence metric (RTD^2) to identify n-grams that contributed to disagreement between base-divergence results derived from n-gram distributions. More specifically, we ranked the RTD values calculated from the ranks of the RTD contributions to divergence results for ground truth and predicted distributions (using our classifiers). Said another way, in cases where n-grams had high RTD^2 values, those n-grams would either be over- or under-emphasized in the data resulting from our classification pipeline when compared with the ground truth.

5.3.5 In-group and out-group prediction

We inferred membership of individual users in in-group or onlooker subreddits using two binary classification models. These models were applied to the entire concatenated comment history of users for a given subreddit. In addition to the data filtering described in section 5.3.3, we removed users whose concatenated comment histories

contained fewer than 10 1-grams. In order to investigate the effect of comment length on classification performance, we created a second training and evaluation data set—referred to as the "threshold" data set—with users whose comment histories contained at least 100 1-grams and who made at least 10 comments on their assigned subreddit. Due to the large class imbalance in most subreddit pairings, we under-sampled the majority class to rebalance the training and testing data sets.

To establish a baseline, we trained a logistic regression model on term frequency-inverse document frequency (TF-IDF) features. For the logistic regression model, we generated TF-IDF features by selecting 1-grams that appeared in at least 10 documents and at most 95% of total documents. We also removed English stopwords before feeding these features to a logistic regression model.

We compared the performance of the logistic regression model with a Longformer-based classifier [18]. The Longformer model uses a sparse attention mechanism to address the quadratic memory scaling of the standard transformers [164]—in our cases allowing for the consideration of longer documents (comment histories). For the Longformer model, we used the default Transformers library [166] implementation of a sequence classifier with a maximum sequence length of 2,048.

5.4 Results

5.4.1 Language classifier

For all subreddit pairs, we found that both language classifiers performed better than random, with some variation along subreddit size and community characteristics, as in

figs. 5.4 and 5.5. The Longformer model performed better in all cases, as indicated by the Matthews correlation coefficient (MCC, see section 1.2.2 for details) in Table 5.1. However, with sufficient data volume, the logistic regression classifier was able to achieve comparable results, especially notable given the reduced model complexity.

For the Longformer model trained and evaluated on NoNewNormal and CovIdiots, we achieved precision and recall values of approximately 0.75 for both classes table 5.5. For the other subreddits, precision and recall values ranged between approximately 0.65 and 0.9 with near parity between the classes. See fig. 5.2 for receiver operator characteristic (ROC) curves for the Longformer model, and see section 1.2.2 for an explanation of this metric.

The logistic regression classifier offered lower performance but relatively similar results with the added benefit of interpretable feature importance scores. In the case of NoNewNormal and CovIdiots, we report feature importance for the logistic regression model in Table 5.3. The feature importance results provide some insights on how bag-of-words models are capturing community-specific language. For instance, “media”, “doomer”, and “trump” are language features highly predictive of the NoNewNormal subreddit accounts. On the other hand, “idiots”, “crocs”, and “5g” are language features highly predictive of the CovIdiots accounts.

5.4.2 Divergence results

Initial observations

We found that RTD identified salient terms when comparing the 1-gram distributions of NoNewNormal and CovIdiots. As seen in Fig. 5.1, we found that terms relating

to specific people and institutions such as “trump”, “fda”, and “fauci” drove RTD contributions from the NoNewNormal distribution. For the same subreddit, we found 1-grams related to vaccines—“vaccine[s]”, “dtp” (Diphtheria-Tetanus-Pertussis), and “npafp” (Non-polio Acute Flaccid Paralysis)—which ranked higher than the opposing subreddit. Finally, some 1-grams related to non-pharmaceutical interventions ranked relatively higher in the NoNewNormal distribution, including “lockdown” and “passport”. From the CovIdiots 1-gram distribution, we saw the eponymous term “covidiot” contributing the greatest to RTD followed by insults such as “stupid” and “karen”—illustrating the insulting critiques that many of the CovIdiots posts level at NoNewNormal.

The RTD results suggest a few characteristics of each subreddit. Both NoNewNormal and CovIdiots discussed prominent topics related to the pandemic—as seen by terms such as “mask”, “vaccine”, and “lockdown” ranking in the top 300 1-grams for each subreddit. The subreddits’ focuses constrast each other with NoNewNormal appearing more focused on discussion that is critical of pandemic interventions and CovIdiots criticizing NoNewNormal (as evidenced by a higher degree of insulting language).

Effect of classifier on divergence results

Overall RTD values were similar for both the ground truth and predicted distributions ($D^{\mathrm{R}}_{1/3} = 0.286$ and 0.274, respectively). In Table 5.2 we present the top 20 1-grams as highlighted by RTD^2 . We saw fluctuations for terms related to internet memes (e.g., “gunga”, “ginga”, and “boo”). In other cases, function words like “he” and “be” are ranked as contributing notably to the RTD^2 results—this may be owing to nuanced

differences in speech patterns between the two communities that are amplified by the classification and RTD^2 results. For some highly topical 1-grams, such "trump", "covidiot", and "influenza", we found shifts in rank limited to an order of magnitude—in these cases the salient 1-grams contributed more to RTD in the classifier-derived data set, likely owing to the bias of the model.

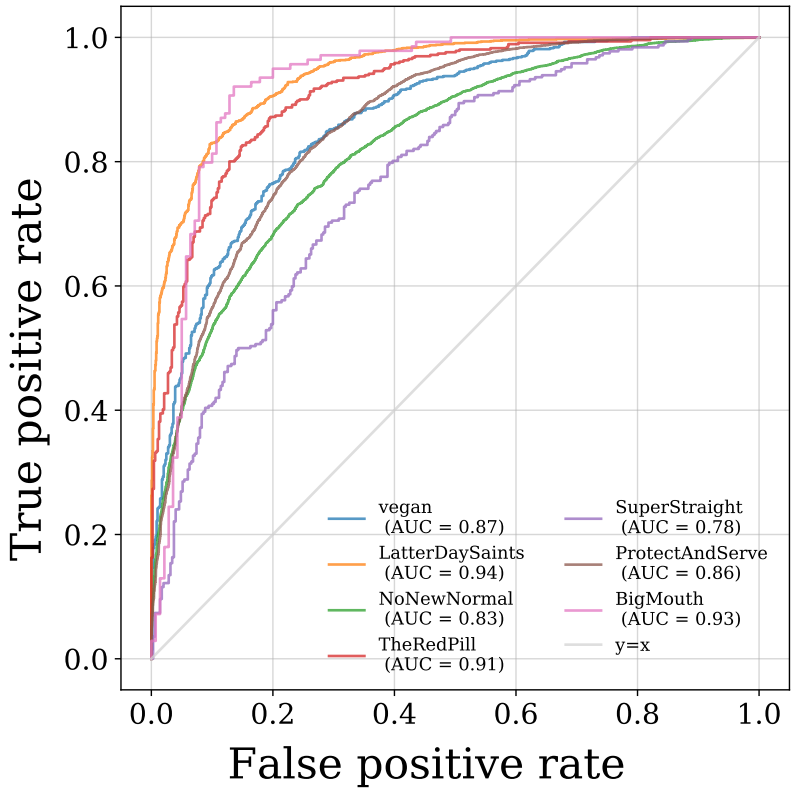


Figure 5.2: Receiver operator characteristic curves for classification models evaluated on the subreddit pairs. For each subreddit pair we trained a binary classifier based on the Longformer language model. The classifier trained on `r/BigMouth` and `r/BanBigMouth` showed the best performance (AUC = 0.93) while our primary case study—`r/NoNewNormal` and `r/CovIdiots`—had an AUC value of 0.83. It is worth noting the variation in sample sizes and as described in Table 5.1.

5.4.3 Accuracy versus user attributes

We expected our classifier to perform better on active users who received praise from a community (as indicated by the voting score on their comments). To confirm this hypothesis, we plotted the likelihood of correctly labeling users that post in NoNewNormal compared to their number of comments in the subreddit, total comment-score, and mean comment-score, shown in fig. 5.3.

Our classifier performed most reliably on users with ten to three hundred comments in the subreddit, and ten to five hundred total karma. Performance decayed for users with over 400 comments, but there were only 520 users in this category out of about 58,000 NoNewNormal users. Anecdotally, this small subset of users engaged in longer and more general discussions, and as a result, used language that is more common and more difficult to classify compared to their less active peers.

To filter out low-activity users, we re-ran our classifier after pruning accounts with less than under 100 one-grams in their comment history or less than 10 total comment in their associated subreddit. This filtering is discussed in section 5.3.5 and labeled "Threshold" in table 5.1 where we present the classification results. The threshold data generally improved the performance of both the logistic regression and Longformer models.

5.5 Discussion

The work outlined here is motivated by the challenge of accurately classifying communities that discuss the same topics but are distinct in their exact views. Further,

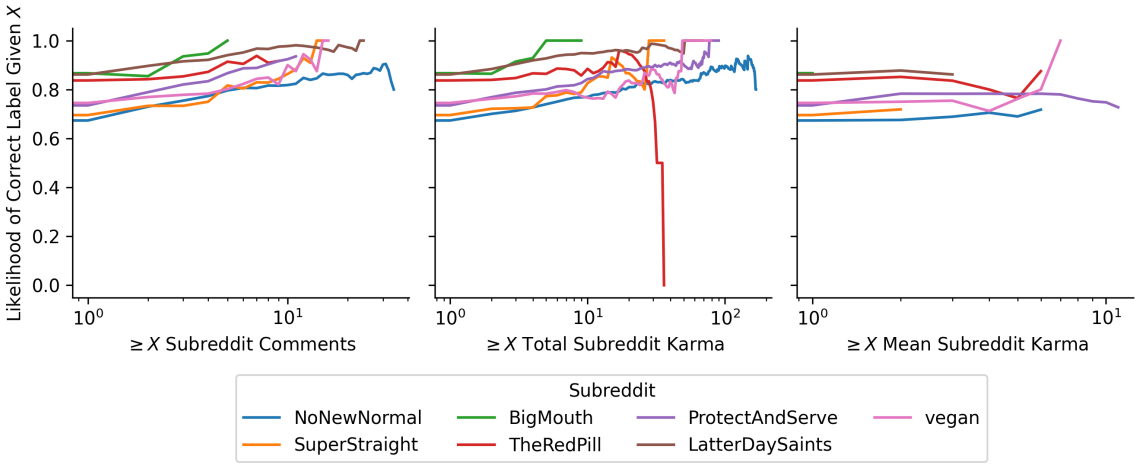


Figure 5.3: Likelihood of correctly labeling users in in-group subreddits by user attributes. From left to right, correct labeling versus user comments in the subreddit, correct labeling versus total karma in the subreddit, and correct labeling versus mean karma in the subreddit. In all cases, the classifier performed poorly with low-activity users, better with moderate activity. We have pruned the 10% of users with the highest attributes from this plot, to improve legibility. An unabridged version of the plot is in the appendix, with a more detailed explanation. Plots include only users that commented in the primary "of" subreddit. Results from base-LR classifier.

we are motivated by the task of identifying these communities in the absence of interaction data that may allow for the construction of a social graph.

Our methodology addresses the challenge of analyzing online conversation around contentious topics where there may be polarized communities that share similar linguistic features. For instance, when studying online discourse around a specific topic one approach to collecting relevant content is anchor wording (selecting posts based on the presence of key words defined by a researcher). In the case of NoNewNormal and CovIdiots, "vaccine", "mask", and "covid" share similar rank values in the 1-gram distributions for each subreddit (55, 37; 24, 28; 51, 58; respectively). A naive anchor-word selection would capture much of the conversation in each of these communities. However, anchor word selection would fail to disambiguate the dramatically differing

Subreddits	MCC				Data set size	
	Base		Threshold		Base	Threshold
	LR	LF	LR	LF		
NoNewNormal v. Covidiots	0.41	0.48	0.57	0.60	44185	6778
TheRedPill v. TheBluePill	0.55	0.65	*	*	4680	402
BigMouth v. BanBigMouth	0.64	0.80	*	*	1394	140
SuperStraight v. SuperStraightPhobic	0.35	0.43	*	*	3310	584
ProtectAndServe v. BadCopNoDonut	0.50	0.55	0.65	0.76	41158	6930
LatterDaySaints v. ExMormon	0.65	0.72	0.80	0.83	15062	4122
vegan v. antivegan	0.49	0.56	0.65	0.72	6896	1692

Table 5.1: Data set size and classification performance for logistic regression (LR) and Longformer (LF) models. Subreddit pairs, primary "of" community first, "onlooking" subreddit second. Matthews correlation coefficient (MCC) refers to performance on the test set. The threshold results refer models trained on a thresholded data set where user comment histories must contain at least 100 1-grams and at least 10 comments. Results excluded due to small sample size are represented with an "*".

views held by the majority of users in each community. This has impacts on down stream analysis such as sentiment analysis, tracking narrative diffusion, and topic modelling.

Considering our main motivation was a problem description and initial demonstration of a classification pipeline, we did not extensively explore model architectures or hyperparameters. We included n-gram order in the initial hyperparameter sweep when developing the logistic-regression pipeline, and results suggested that 1-grams were most effective. However, including higher order n-grams is still worth exploring more in-depth, and may have benefits for model interpretabilility and down stream results (e.g., feature importance). Further, we selected the word-embedding model

1-gram	RTD^2 Rank	RTD rank (pred.)	RTD rank (actual)
he	1	11.0	446.0
be	2	4285.0	19.0
vaccin	3	7.0	104.0
thi	4	143.0	8.0
nyt	5	15.0	459.0
they	6	27.0	3414.5
diffrent	7	42.5	17076.0
ginga	8	73.5	9.0
gunga	9	24.0	5.0
shill	10	103.0	13.0
titer	11	11026.0	59.5
boo	12	2.0	1.0
covidiot	12	1.0	2.0
sham	14	52.0	4253.0
voluntari	15	53.0	4420.5
influenza	16	14.0	103.0
purg	17	1694.5	44.0
postul	18	16.0	123.0
trump	19	8.0	3.0
dui	20	51.0	1956.0

Table 5.2: Rank-turbulence divergence (RTD) of divergence results from actual and predicted 1-gram distributions. As a divergence-of-divergences measurement, RTD^2, shows disagreement between the divergence results derived from 1-gram distributions of generated with ground truth labels and the distribution generated with our classification pipeline. Highly ranked RTD^2 values highlight the 1-grams that have the greatest difference in rank of contribution to the divergence results for each pairing. For instance, "trump" is the 1-gram with the 3^{rd} highest contribution in ground-truth data, whereas the 1-gram is ranked 8^{th} in the classifier-generated data. We stemmed the 1-grams prior to calculation of divergence results.

(the Longformer) based mainly on considerations related to maximum sequence length and preliminary performance observations. Additional word-embedding models could be considered—choosing models trained on more recent and/or domain specific data may be especially helpful.

As in stance detection [7], there are several limitations to the methodology we

present. First, our data set covers a limited time frame, and past work has demonstrated that models which are trained on old data sets may perform relatively poorly when fed new data [6, 7]. Additionally, our methodology does not account for the fact that users may change opinions throughout time. For example, a user may initially be a member of a group, but a shift in opinion may cause the user to leave the group but still engage in discussion about said group. Lastly, our classifier is only trained on English posts, and we cannot guarantee the same level of performance across languages.

5.6 Conclusion

In the present study, we frame the research challenge of classifying in-groups and onlookers based on the linguistic features of social media posts. The classification task is made difficult by the significant intersection of terms shared between the two communities, which may confound classification attempts. We collect a data set of seven (7) subreddit pairs that match the in-group and onlooker-group criteria, focusing our efforts on a case study of pro- and anti-COVID mitigation communities. These subreddits provide an appealing proving ground for group identification tasks, because subreddit participation acts as a noisy label in lieu of ground truth for group identity. We identify salient 1-grams that differentiate each communities' language distributions. Using the full collection of subreddit pairs, we train two classifiers to assign users to communities based on their posts. We demonstrate the feasibility of the classification scheme with these results. In most cases, our classifier recovers 70% or more of a community's users. From these results, we show how our initial

language distribution divergence results may be affected by using data labelled by our classifier. In the case of the COVID subreddits, the true and classifier-generated distributions are qualitatively similar, identifying notable 1-grams in each case. We hope the research questions and combined set of results is motivating for future work that leverages training generalizable classifiers on labelled community data that can then be used in a variety of settings.

5.7 Future Work

We present a first attempt at in-group classification based on contextual language use, in a challenging environment where both the in-group and onlookers discuss many of the same topics. We believe that classifiers in this domain have important applications for cross-platform group detection, where more reliable labels like consistent usernames and network interactions are unavailable. More powerful classifiers may account for additional text features, including user sentiment, shared topics, stance towards those topics, and language style. Longer time-span studies should be wary of semantic drift over time [142], as well as more specific changes in group language and stance on topics. Models of community language style [157] could also help identify communities across platforms, as long as platform-specific language style features are identified and controlled for.

Appendix

Subreddit Corpus Sizes

Table 5.4 indicates the size of each subreddit, in terms of user count and comment count, after pruning bots and low-karma users as specified in our methodology. It also includes the mean karma (comment score) for remaining comments in each subreddit corpus.

Comparison of Subreddit Activity

If subreddits in a pair have dramatically different activity levels, such as much longer comments in one subreddit than another, these differences in writing style may correlate with classification difficulty. figs. 5.4 and 5.5 show cumulative distributions of comment length and comment count per user, respectively, to illustrate which subreddits are closer in behavior than others.

Uniquely Identifying Words

table 5.3 shows the words that most strongly correlate with membership in NoNewNormal and CovIdiots.

Labeled Language versus Predicted Language

fig. 5.1 shows word use divergence between NoNewNormaland CovIdiotsusing all comments from users in each subreddit. For comparison, fig. 5.7 shows the same word use

divergence based only on users our classifier predicted as members of each subreddit.

Classifier performance metrics

table 5.5 shows F1 scores and precision values for the logistic regression and longformer model.

Classifier Accuracy versus User Attributes

Our classifier performs best on accounts with above 10 comments and a minimum comment-karma threshold. However, the classifier cannot reliably label every user in the tail of the distribution. This leads to a misleading visualization, conflating the low-density of users that have high comment counts or karma scores with classifier performance. Therefore, we did not include the tail of each performance graph in fig. 5.3. For posterity, we have included an unabridged version of the graph that includes these misleading tails, in fig. 5.6.

r/NoNewNormal	r/CovIdiots
media	covidiots
emails	covidiot
questioning	retard
lockdown	cunt
jab	nnn
power	report
restrictions	idiot
narrative	deniers
woke	idiots
yall	idiocy
guys	crocs
passport	ugh
msm	5g
subreddit	selection
dystopian	wedding
sheep	frustrating
doomer	fox
doomers	hoax
sub	beard
trump	department

Table 5.3: Feature importance for logistic regression classifier trained on r/NoNewNormal and r/CovIdiots. The two columns correspond to the text features that are most strongly predictive of each subreddit.

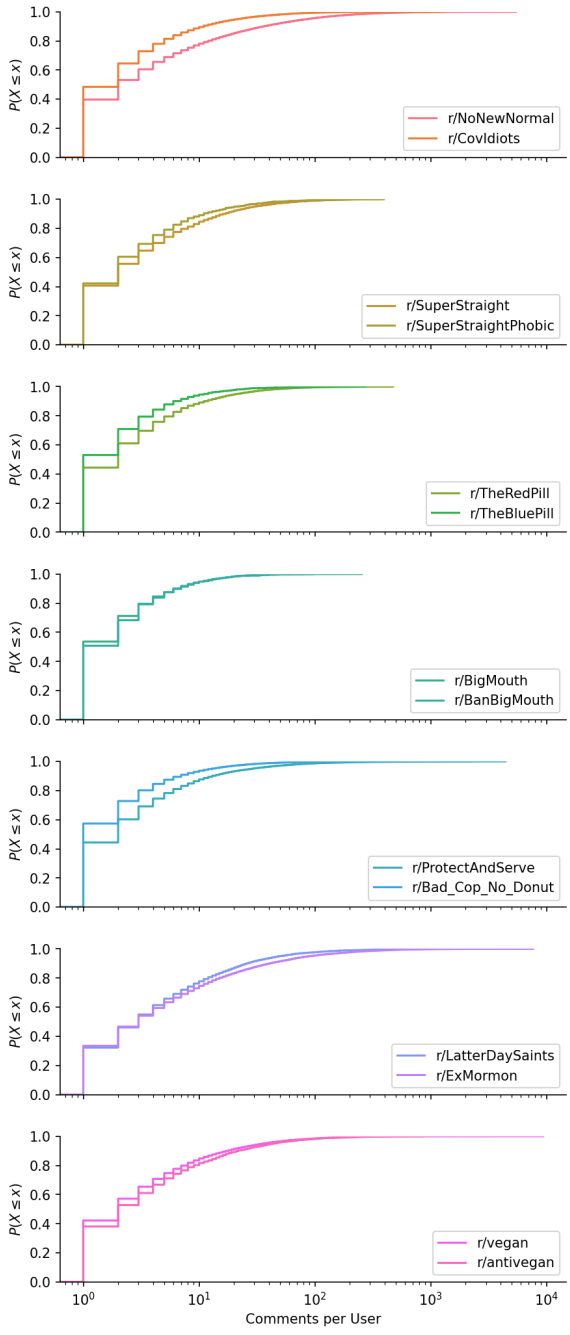


Figure 5.4: Cumulative distribution of comments made by each user in each examined subreddit pair. Distribution taken after filtering.

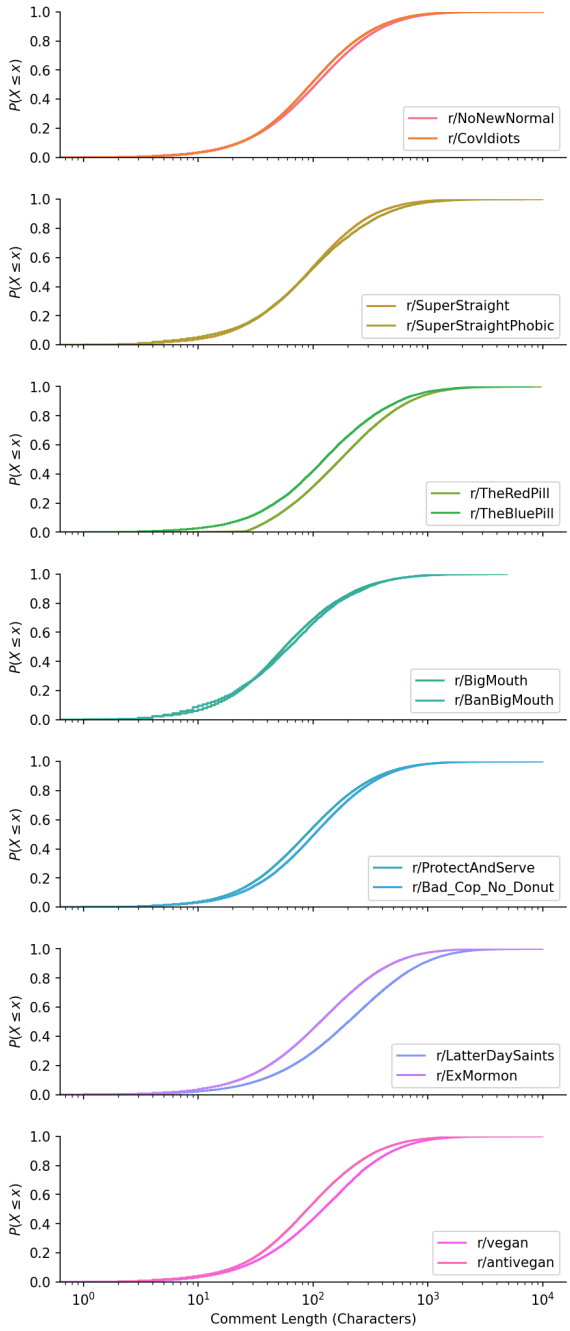


Figure 5.5: Cumulative distribution of comment length in each examined subreddit pair. Distribution taken after filtering.

Subreddit	Users	Comments	Mean Karma
r/NoNewNormal	57966	1245398	4.743
r/CovIdiots	28427	174056	4.119
r/TheRedPill	10149	59388	3.608
r/TheBluePill	2744	9616	4.716
r/BigMouth	6252	19904	1.895
r/BanBigMouth	981	3226	1.359
r/SuperStraight	5914	46491	2.686
r/SuperStraightPhobic	1897	11498	1.449
r/ProtectAndServe	25096	241328	7.484
r/Bad_Cop_No_Donut	77288	314933	5.898
r/LatterDaySaints	9130	131055	2.498
r/ExMormon	35672	852607	3.440
r/vegan	62544	622069	4.908
r/antivegan	4492	47738	3.878

Table 5.4: Users and comments in each subreddit, after filtering out bots and low-karma users

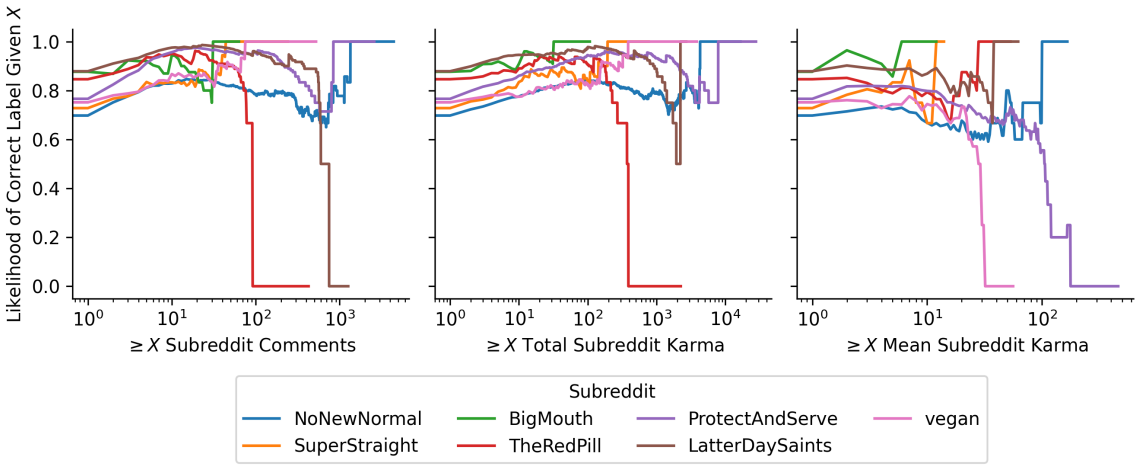


Figure 5.6: Likelihood of correctly labeling users in in-group subreddits by user attributes. This is the unabridged version of fig. 5.3, including unstable long-tail behavior when classifying the small minority of high-activity accounts.

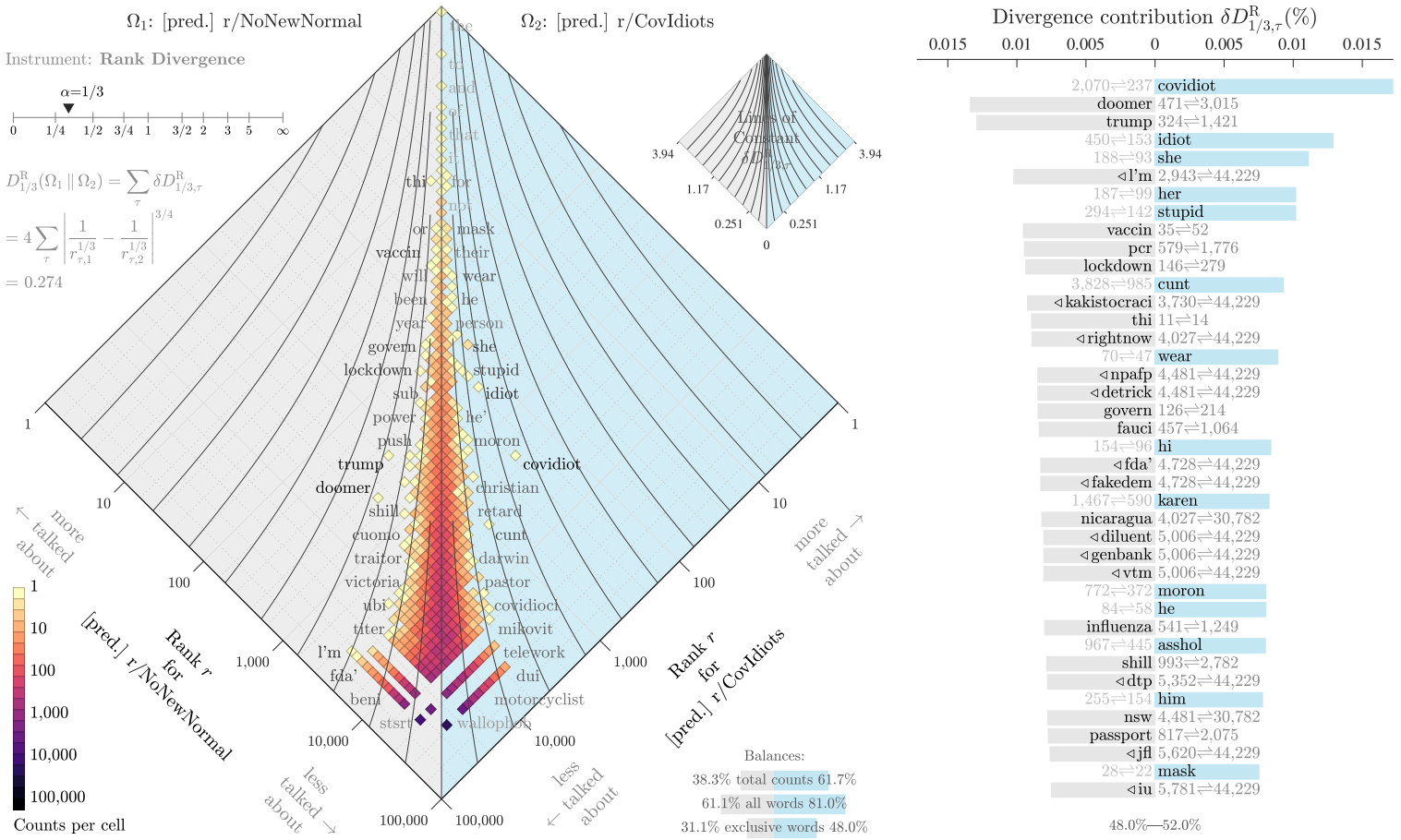


Figure 5.7: An allotaxonograph [41] showing the 1-gram rank distributions of predicted users of `r/NoNewNormal` and `r/CovIdiots` using our classifier to assign membership. See Fig. 5.1 for allotaxonograph of actual users. The central diamond shaped plot shows a rank-rank histogram for 1-grams appearing in each subreddit. The horizontal bar chart on the right show the individual contribution of each 1-gram to the overall rank-turbulence divergence value ($D^{R}_{1/3}$). The 3 bars under "Balances" represent the total volume of 1-gram occurring in each subreddit, the percentage of all unique words we see in each subreddit, and the percentage of words that we see in a subreddit that are unique to that subreddit.

Subreddits	F1				Precision				Data set size	
	Base		Threshold		Base		Threshold		Base	Threshold
	LR	LF	LR	LF	LR	LF	LR	LF		
NoNewNormal v. Covidiots	0.71	0.74	0.83	0.80	0.71	0.74	0.83	0.80	44185	6778
TheRedPill v. TheBluePill	0.79	0.84	*	*	0.84		*	*	4680	402
BigMouth v. BanBigMouth	0.80	0.88	*	*	0.80	0.88	*	*	1394	140
SuperStraight v. SuperStraightPhobic	0.67	0.69	*	*	0.67	0.69	*	*	3310	584
ProtectAndServe v. BadCopNoDonut	0.75	0.78	0.90	0.88	0.75	0.78	0.90	0.88	41158	6930
LatterDaySaints v. ExMormon	0.83	0.86	0.95	0.91	0.83	0.86	0.95	0.91	15062	4122
vegan v. antivegan	0.75	0.78	0.88	0.86	0.75	0.78	0.88	0.86	6896	1692

Table 5.5: Data set size and classification performance for logistic regression (LR) and Longformer (LF) models. Subreddit pairs, primary "of" community first, "onlooking" subreddit second. F1 scores and precision values are calculated using weighted average for the balanced data sets. F1, precision, and recall (not shown) values were all approximately equal for specific models and subreddit pairs in our experiments—partially owing to the balanced datasets. The threshold results refer models trained on a thresholded data set where user comment histories must contain at least 100 1-grams and at least 10 comments. Results excluded due to small sample size are represented with an "*".

Chapter 6

Discussion

6.1 Key Findings and Implications

My research examines the interplay between platform design and group behavior at a variety of scales, including how users are impacted by the operational rules of a platform, how communities are impacted by platform social policy changes, how communities influence one another within platform boundaries, and how communities transcend platform boundaries.

My findings reflect the heterogeneity of human experience. Differences between development practices on GitHub and the Penumbra are visible only in large aggregate. Even within our limited sample of fifteen banned subreddits, group response to bans varied widely. When measuring inter-group influence and platform centralization, Voat is entirely unique in how its largest two communities by orders of magnitude have no population overlap with the rest of the platform.

Despite that variance, we do find some emergent patterns. Projects developed off of GitHub tend to have more collaborators, and are maintained for longer with more

consistency. GitHub's extremely public nature lends itself both to "portfolio projects" archived with no intention of ongoing development, and to "drive-by contributions," where a developer adds a new feature or a bug-fix to a project, then departs without any further engagement. These differences do not appear to be driven by the technical features afforded by GitHub, but by the collective rules of how users interact within GitHub's social context.

Regarding deplatforming, our results suggest that community bans are most effective against groups with a clear social identity. Users from communities banned for casual racism and "dark humor" can find a variety of similar groups to participate in. The communities whose member activity and plummeted most were extreme right-wing groups, and the anti-trans "gendercritical" community. All of these groups have highly-specific in-group vocabulary and belief systems that are widely unpopular and difficult to integrate even into other subreddits. Importantly, a drop in Reddit activity does not suggest that these communities become inactive, but is as likely to indicate that they have moved off-platform, often to an alt-tech platform more welcoming to them.

I have argued that a community's influence on its peers is a function of both its size and topological role within a platform. Platforms with a heavily skewed community size distribution are not necessarily centralized, as illustrated by Voat, where the QAnon and 8chan communities established a large presence without engaging with the rest of the site. Studying only the largest communities on such a platform, or even randomly sampling activity across the whole platform, will give a biased perspective that over-represents the largest group's role.

My analysis of inter-community influence also shows the shortcomings of some

decentralization efforts. Mastodon was popularized as a "decentralized" alternative to Twitter that incorporated the "forking" culture of open-source. Rather than being run by a single company with one global social policy, Mastodon servers each have their own administration and policies, and users can secede by creating their own servers and bringing their accounts and followers with them. However, Mastodon was more centralized than any other platform we examined: not only did the three largest servers encompass half the platform's population, but they were deeply interconnected with the rest of the platform, so the social policies set on those three servers have a profound impact on what content is seen by users across Mastodon. As with GitHub, this has less to do with the operational rules of the platform than the collective rules. Mastodon's features for migrating accounts and facilitating inter-server follows do ostensibly enable decentralization. However, two social pressures are in conflict with these goals. First, social media that revolves around following users (as with X, Mastodon, Instagram) lends itself to rich-get-richer dynamics where a small minority of users receive an exceptional amount of engagement and follows. On Mastodon, this means the servers with early popular accounts receive more users and more inter-server follows. Therefore, Mastodon tends towards a small number of much larger instances well-connected to the rest of the platform. Additionally, while servers can nominally have independent social policies, many server administrators will not "federate" (allow their servers to exchange content) with servers that do not have at least similar policies. This applies positive social pressure - servers with no content moderation that implicitly permit hate speech and harassment tend to be isolated from the rest of Mastodon - but it also means that the platform tends towards a global mono-culture social policy primarily dictated by the administrators of the

largest servers.

My work on inter-platform community migration is in its early stages. My efforts so far have focused on developing a "group linguistic fingerprint," whereby instead of trying to match users across two different platforms by a metric like username similarity, we can instead argue that two groups discuss the same subjects with the same language used in contextually the same way, and so are likely the same community. My hope is that this community fingerprint will allow researchers to study the behavior of multi-platform communities, and the effects of deplatforming and platform migration, at scale without deanonymization and tracking individual users. So far we have demonstrated that classifiers can readily distinguish between members of an in-group and people discussing the in-group based on linguistic features. Ongoing work suggests that while classifiers lose efficacy over time as conversation topics and vocabulary diverge from older training data, it may be possible to compensate for this degradation by identifying language that is both consistent throughout a community's timeline and is uniquely identifying compared to other communities.

6.2 Limitations and Caveats

Most of my chapters feature observational experiments, where I am monitoring a community without careful control. For example, chapter 2 contrasts open source projects on and off of GitHub to infer GitHub's influence, where a controlled experiment might force projects to migrate from GitHub to the Penumbra or vice-versa, observing how a change in operational and collective rules impacts project development. Instead, developer communities self-select whether to host their projects on or

off of GitHub, and may choose not to host on GitHub for ideological or bureaucratic reasons rather than due to technical or social affordances. For example, we observe several large open source projects on the Penumbra, who may be resistant to hosting their code on Microsoft-controlled infrastructure due to Microsoft's historical hostility towards open-source projects and open standards. We identify a number of universities hosting their own GitLab infrastructure for students and faculty, who may be motivated to self-host by data control and archival requirements in grant funding or easier integration with other university services. While we control for academic-status in our analysis, we cannot disambiguate between the effect that GitHub has on project development, and a predisposition for certain development practices among the kinds of projects that choose not to host on GitHub.

Similarly, in chapter 3 we do not examine how *arbitrary* subreddits respond to being deplatformed, but specifically how fifteen of the largest subreddits banned in 2020 responded. This is a small sample size, but it is also a very selective sample. These subreddits were banned for hate, harassment, and incitement of violence, and were predominantly politically far-right. Their response to deplatforming may not be representative of groups banned for other reasons, such as illicit drug markets, financial crime, or sex work. This remains a useful sample for understanding other online hate and extremism, but its generalizability should not be overstated.

I make a largely implicit assumption that accounts correlate with human users. While individuals may have more than one account on a platform, it would confound results if, for example, most of the top accounts from a banned subreddit were all operated by a single human. We do not typically have access to information from platforms that would alleviate these concerns, such as access logs showing the IP

addresses accounts posted from. However, the platforms *do* have access to such information, and often have policies about exploiting the collective rules of the platform through "sockpuppet" accounts. For example, Reddit permits users to have multiple accounts for different contexts, but forbids anyone from engaging in vote-manipulation by using several accounts to artificially increase the score of one post and deflate the scores of surrounding posts.

6.3 Future Work Left Undone

My work engages with how the affordances offered by online platforms, and the social policies instated on them, influence online group behavior. I have examined patterns of behavior on several platforms as case studies, and have drawn comparisons between the structure of different platforms. What I have not begun to address are many open questions related to governance of platforms, and inter-platform dynamics.

On governance, we have some understanding of how individual social policies, such as Reddit community bans (chapter 3) or YouTube bans of COVID-19 misinformation [122], may influence group behavior. However, we have less understanding of how governance structure itself influences a community. For example, platforms like Reddit and Discord are defined by shared governance; the platforms themselves are run by private corporations that can unilaterally set content policies and change platform functionality, but communities within those platforms ("subreddits" on Reddit, "servers" on Discord) are governed by volunteer moderators who are members of those communities. These moderators are given both authority and technical tools to set and enforce community guidelines beyond what is permissible on the platform,

and have a formative influence on community culture and normative behavior.

There has been considerable qualitative work on how to categorize online platforms by governance style and social affordances [48, 131, 184]. In particular, there is a line of research on governance within open source software projects, because these communities are more intentionally organized than social media with explicit goals around coordination, software development, and long-term maintenance. Pioneering work in this space categorized open source projects as centrally-designed and privately developed "cathedrals" or community-contributed and organically-design "bazaars" [133], and further work broke down governance styles into an axis from a "benevolent dictator" model where a single individual is responsible for project decisions to a range of "community consensus" models where decisions are made either by group vote or by direct-action and community veto [89]. I believe that there is a need to complement these frameworks with quantitative analysis of observed large-scale behavior in many different kinds of online social communities, to improve insight in to how to pick technical and social affordances to foster a desired community outcome.

There are many studies focusing on behavior and content within one platform. Reductionism trains scientists to reduce phenomenon to the smallest observable unit, controlling as many variables as possible to simplify problems. Unfortunately for this framework, online platforms do not exist independently from one another, but are part of a shared online ecosystem. Users flow from one platform to another with ease, often maintaining footholds on many platforms at once, carrying information and screenshots between them. Social policies instated on one platform will therefore influence adjacent platforms, by drawing users to their platform and creating new links, or by driving users away and changing the populations on other platforms, or

by inhibiting certain behavior and information within their own walls that will no longer be copied beyond their borders.

Well-developed research on online community behavior will necessarily engage with multiple platforms at once. We already have many case studies that fit this description, such as studies of link-sharing where users on one platform post links to a second platform [31], or studies of "raids" wherein hostile users from one platform harass users on another [74]. We also have limited insight into community migration between platforms, when those migrations are well-established, such as the founding of `thedonald.win` after Reddit administration quarantined `r/the_donald` [78]. It is my hope that with approaches like those I outlined in chapter 5 we can systematically study the migration of communities between platforms.

6.4 Ethics and the Future of our Field

An existential crisis facing studies of online group behavior is a lack of accessible information. This comes from a combination of corporate policy change, and community movement into more "opaque" platforms. In early 2023, Twitter cut off academic access to public data (or made it prohibitively expensive to use their research API) among a number of sweeping changes made by new owner Elon Musk. In April of that year, Reddit announced they would close most of their free API access, intending to charge money for companies to use Reddit posts and comments as large language model training data. Meta has long had an antagonistic relationship with researchers unless they collaborate directly with Meta's research teams within restrictive data sharing agreements. In addition to corporate policy decisions, communities may mi-

grate to platforms that do not offer any means to browse information "publicly." For example, while you can browse a subreddit without being logged in, you cannot read the messages in a Discord server or Telegram channel without explicitly, and typically visibly, joining the chatroom. This can be an uncomfortable requirement for academic researchers, who often use an analogy between posting comments on the public Internet and publishing in a newspaper to justify that their studies are passive, do not interact with human subjects, and therefore do not require IRB approval.

There are no easy answers to how our field can address a dearth of public community behavioral data. In some contexts, it may be justifiable to break platform terms of service, scraping content for analysis without the approval of the administration. I believe this is particularly acceptable when studying online hate or similar hostile behavior, where there is a moral argument that understanding and inhibiting harmful behavior is more important than abiding by a platform's request that no automated tools interact with their website. Similarly, researchers may need to grow more comfortable with using bots to enter chatrooms to gather data, perhaps seeking IRB approval where necessary. Occasionally, we obtain insight into platforms through hacked, leaked, or scraped datasets. For example, shortly after the January 6$^{\text{th}}$ Capitol riots, individuals scraped all videos from Parler to preserve evidence posted during or after the event [119]. The following month, individuals hacked Gab, the alt-tech Twitter-like service, and leaked that dataset to Distributed Denial of Secrets, who made it available to academics and journalists [154]. As we lose more "legitimate" access to online data through public APIs, researchers may have little choice but to adopt a broader set of sources if we want insight into digital society.

6.5 Closing Remarks

Clark and Chalmers introduced the *extended mind theory* [10], which posits that our minds include cognitive processes beyond our brains, beyond our nervous systems or bodies, encompassing information storage and processing tools such as diaries, calculators, and computers. Following this logic, social media can be thought of as a shared cognitive process taking place across many minds. The operational and collective rules of a platform both facilitate and constrain what interactions can occur within it, and therefore what kind of shared cognition is possible. The design of collaborative online platforms is no less than encouraging a vision of society, and as such this design process should be driven by hope and idealism. Much of my research has focused on the negative side of online human interaction, including hate and radicalization, and the centralization of authority into the hands of the few. Nevertheless, I would like to frame my work as a small contribution towards a kinder society, with more collective and egalitarian ideals.

Bibliography

[1] Gavin Abercrombie and Riza Theresa Batista-Navarro. "Identifying opinion-topics and polarity of parliamentary debate motions". In: *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*. Association for Computational Linguistics. 2018.

[2] Jean-François Abramatic, Roberto Di Cosmo, and Stefano Zacchiroli. "Building the universal archive of source code". In: *Communications of the ACM* 61.10 (2018), pp. 29–31.

[3] Hind S Alatawi, Areej M Alhothali, and Kawthar M Moria. "Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT". In: *IEEE Access* 9 (2021), pp. 106363–106374.

[4] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of modern physics* 74.1 (2002), p. 47.

[5] Réka Albert, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks". In: *nature* 406.6794 (2000), pp. 378–382.

[6] Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. "Opinions are made to be changed: Temporally adaptive stance classification". In: *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*. 2021, pp. 27–32.

[7] Rabab Alkhalifa and Arkaitz Zubiaga. "Capturing Stance Dynamics in Social Media: Open Challenges and Research Directions". In: *arXiv preprint arXiv:2109.00475* (2021).

[8] Abhinav Anand and Jalaj Pathak. "The role of Reddit in the GameStop short squeeze". In: *Economics Letters* 211 (2022), p. 110249.

[9] Pranav Anand et al. "Cats rule and dogs drool!: Classifying stance in online debate". In: *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*. 2011, pp. 1–9.

[10] Clark Andy and Chalmers David. "The extended mind". In: *Analysis* 58.1 (1998), pp. 7–19.

[11] Ziqiao Ao, Gergely Horvath, and Luyao Zhang. "Are decentralized finance really decentralized? A social network analysis of the Aave protocol on the Ethereum blockchain". In: *arXiv preprint arXiv:2206.08401* (2022).

[12] R. Armitage. "Online 'anti-vax' campaigns and COVID-19: censorship is not the solution". In: *Public Health* 190 (Jan. 2021), e29–e30. ISSN: 00333506. DOI: 10.1016/j.puhe.2020.12.005.

[13] Isabelle Augenstein et al. "Stance detection with bidirectional conditional encoding". In: *arXiv preprint arXiv:1606.05464* (2016).

[14] Albert-László Barabási and Eric Bonabeau. "Scale-free networks". In: *Scientific american* 288.5 (2003), pp. 60–69.

[15] Jason Baumgartner et al. "The pushshift Reddit dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 2020, pp. 830–839.

[16] Matthew D. Beckman et al. "Implementing Version Control With Git and GitHub as a Learning Objective in Statistics and Data Science Courses". In: *Journal of Statistics and Data Science Education* 29.sup1 (2021), S132–S144.

[17] Andrew Beers et al. *The Firestarting Troll, and Designing for Abusability.* 2021.

[18] Iz Beltagy, Matthew E Peters, and Arman Cohan. "Longformer: The long-document transformer". In: *arXiv preprint arXiv:2004.05150* (2020).

[19] Sumit Bhatia and P Deepak. "Topic-Specific Sentiment Analysis Can Help Identify Political Ideology". In: *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2018, pp. 79–84.

[20] Lindsay Blackwell et al. "Classification and its consequences for online harassment: Design insights from heartmob". In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), pp. 1–19.

[21] Phillip Bonacich. "Power and centrality: A family of measures". In: *American Journal of Sociology* 92.5 (1987), pp. 1170–1182.

[22] Cristina Bosco et al. "Overview of the EVALITA 2018 hate speech detection task". In: *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. Vol. 2263. CEUR. 2018.

[23] Alexandre Bovet and Hernán A. Makse. "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature Communications* 10.1 (Dec. 2019), p. 7. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07761-2.

[24] Carter T Butts. “Revisiting the foundations of network analysis”. In: *science* 325.5939 (2009), pp. 414–416.

[25] Amanda Casari et al. “Open source ecosystems need equitable credit across contributions”. In: *Nature Computational Science* 1.1 (2021), pp. 2–2.

[26] Dorota Celińska. “Coding together in a social network: collaboration among GitHub users”. In: *Proceedings of the 9th International Conference on Social Media and Society*. 2018, pp. 31–40.

[27] Eshwar Chandrasekharan et al. “You can’t stay here: The efficacy of Reddit’s 2015 ban examined through hate speech”. In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), pp. 1–22.

[28] Hongxu Chen et al. “Multi-level Graph Convolutional Networks for Cross-platform Anchor Link Prediction”. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, Aug. 2020, pp. 1503–1511. ISBN: 978-1-4503-7998-4. DOI: 10.1145/3394486.3403201. URL: https://dl.acm.org/doi/10.1145/3394486.3403201.

[29] Hsin-liang Chen and Yin Zhang. “Functionality analysis of an open source repository system: current practices and implications”. In: *The Journal of Academic Librarianship* 40.6 (2014), pp. 558–564.

[30] Davide Chicco and Giuseppe Jurman. “The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification”. In: *BioData Mining* 16.1 (2023), p. 4.

[31] Matthew Childs et al. “Characterizing YouTube and BitChute Content and Mobilizers During US Election Fraud Discussions on Twitter”. In: *14th ACM Web Science Conference 2022*. 2022, pp. 250–259.

[32] Samridhi Shree Choudhary et al. “Modeling Coordination and Productivity in Open-Source GitHub Projects”. In: *Carnegie-Mellon Univ. Inst of Software Research International, Tech. Rep* (2018), CMU-ISR-18–101.

[33] Fan Chung and Linyuan Lu. “The average distances in random graphs with given expected degrees”. In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pp. 15879–15882.

[34] Curtis Clifton, Lisa C. Kaczmarczyk, and Michael Mrozek. “Subverting the Fundamentals Sequence: Using Version Control to Enhance Course Management”. In: *SIGCSE Bull.* 39.1 (Mar. 2007), pp. 86–90.

[35] Hugo Coll et al. “Free software and open source applications in higher education”. In: *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*. 5. WSEAS. 2008.

[36] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[37] Harald Cramér. *Mathematical methods of statistics*. Vol. 26. Princeton university press, 1999.

[38] Laura Dabbish et al. "Social coding in GitHub: transparency and collaboration in an open software repository". In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 2012, pp. 1277–1286.

[39] Brittany I Davidson et al. "Social Media APIs: A Quiet Threat to the Advancement of Science". In: *PsyArXiv* (2023).

[40] Gabriele Di Bona et al. *The decentralized evolution of decentralization across fields: from Governance to Blockchain*. Number: arXiv:2207.14260. July 28, 2022. arXiv: 2207.14260[physics]. URL: http://arxiv.org/abs/2207.14260 (visited on 08/08/2022).

[41] Peter Sheridan Dodds et al. "Allotaxonometry and rank-turbulence divergence: A universal instrument for comparing complex systems". In: *arXiv preprint arXiv:2002.09770* (2020).

[42] Joan Donovan, Becca Lewis, and Brian Friedberg. "Parallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech". en. In: (2019). DOI: 10.25969/MEDIAREP/12374. URL: https://mediarep.org/handle/doc/13283.

[43] Mohsen Dorodchi and Nasrin Dehbozorgi. "Utilizing open source software in teaching practice-based software engineering courses". In: *2016 IEEE Frontiers in Education Conference (FIE)*. 2016, pp. 1–5.

[44] Paul Erdős, Alfréd Rényi, et al. "On the evolution of random graphs". In: *Publ. Math. Inst. Hung. Acad. Sci* 5.1 (1960), pp. 17–60.

[45] Jenny Fan and Amy X. Zhang. "Digital Juries: A Civics-Oriented Approach to Platform Governance". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20: CHI Conference on Human Factors in Computing Systems. Honolulu HI USA: ACM, Apr. 21, 2020, pp. 1–14. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376293. URL: https://dl.acm.org/doi/10.1145/3313831.3376293 (visited on 02/05/2022).

[46] Joseph Feliciano, Margaret-Anne Storey, and Alexey Zagalsky. "Student experiences using GitHub in software engineering courses: a case study". In: *2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C)*. IEEE. 2016, pp. 422–431.

[47] Linton C. Freeman. "Centrality in social networks conceptual clarification". In: *Social Networks* 1.3 (Jan. 1978), pp. 215–239. ISSN: 03788733. DOI: 10.1016/0378-8733(78)90021-7. URL: https://linkinghub.elsevier.com/retrieve/pii/0378873378900217 (visited on 08/08/2022).

[48] Seth Frey, P. M. Krafft, and Brian C. Keegan. “"This Place Does What It Was Built For": Designing Digital Institutions for Participatory Change”. In: *Proceedings of the ACM on Human-Computer Interaction* 3 (CSCW Nov. 7, 2019), pp. 1–31. ISSN: 2573-0142. DOI: 10.1145/3359134. URL: https://dl.acm.org/doi/10.1145/3359134 (visited on 02/04/2022).

[49] Seth Frey, PM Krafft, and Brian C Keegan. ““This Place Does What It Was Built For" Designing Digital Institutions for Participatory Change”. In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–31.

[50] Mei Fukuda, Kazuyuki Shudo, and Hiroki Sayama. *Detecting and Forecasting Local Collective Sentiment Using Emojis.* 2022.

[51] Ryan J Gallagher et al. “Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts”. In: *EPJ Data Science* 10.1 (2021), p. 4.

[52] Joshua Garland et al. “Countering hate on social media: Large scale classification of hate and counter speech”. In: *ACL Workshop on Online Abuse and Harms* (2020), pp. 102–112.

[53] Ona de Gibert et al. “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 11–20.

[54] Ona de Gibert et al. “Hate Speech Dataset from a White Supremacy Forum”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018, pp. 11–20.

[55] James J Gibson. “The theory of affordances”. In: *Hilldale, USA* 1.2 (1977), pp. 67–82.

[56] Denver Gingerich and Bradley M. Kuhn. *Give Up GitHub: The Time Has Come!* June 30, 2022. URL: https://sfconservancy.org/blog/2022/jun/30/give-up-github-launch/ (visited on 07/19/2023).

[57] GitHub. *New year, new GitHub: Announcing unlimited free private repos and unified Enterprise offering*. https://github.blog/2019-01-07-new-year-new-github/. Accessed: 2021-06-14. 2019.

[58] GitHub. *The 2020 State of the Octoverse*. https://octoverse.github.com/. Accessed: 2021-06-14. 2020.

[59] Christoph Gote, Ingo Scholtes, and Frank Schweitzer. “git2net: Mining time-stamped co-editing networks from large git repositories”. In: *Proceedings of the 16th International Conference on Mining Software Repositories*. IEEE Press. 2019, pp. 433–444.

[60] Christoph Gote and Christian Zingg. “gambit–An Open Source Name Disambiguation Tool for Version Control Systems”. In: *Proceedings of the 18th International Conference on Mining Software Repositories* (2021).

[61] Camille Grange. “The Generativity of Social Media: Opportunities, Challenges, and Guidelines for Conducting Experimental Research”. In: *International Journal of Human–Computer Interaction* 34.10 (Oct. 2018), pp. 943–959. ISSN: 1044-7318, 1532-7590. DOI: 10.1080/10447318.2018.1471573.

[62] Ilya Grigorik. *The GitHub Archive*. https://githubarchive.org. 2012.

[63] Lassi Haaranen and Teemu Lehtinen. “Teaching git on the side: Version control system as a course platform”. In: *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education*. 2015, pp. 87–92.

[64] Hussam Habib et al. “To Act or React: Investigating Proactive Strategies For Online Community Moderation”. In: *arXiv preprint arXiv:1906.11932* (2019).

[65] Margeret Hall et al. “Editorial of the Special Issue on Following User Pathways: Key Contributions and Future Directions in Cross-Platform Social Media Research”. In: *International Journal of Human–Computer Interaction* 34.10 (Oct. 2018), pp. 895–912. ISSN: 1044-7318, 1532-7590. DOI: 10.1080/10447318.2018.1471575.

[66] Momchil Hardalov et al. “A survey on stance detection for mis-and disinformation identification”. In: *arXiv preprint arXiv:2103.00242* (2021).

[67] Areej Al-Hassan and Hmood Al-Dossari. “Detection of hate speech in social networks: a survey on multilingual corpus”. In: *6th International Conference on Computer Science and Information Technology*. 2019.

[68] K Hazel Kwon and Chun Shao. “Communicative Constitution of Illicit Online Trade Collectives: An Exploration of Darkweb Market Subreddits”. In: *International Conference on Social Media and Society*. 2020, pp. 65–72.

[69] Bing He et al. “Racism is a Virus: Anti-Asian Hate and Counterspeech in Social Media during the COVID-19 Crisis”. In: *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’21. Virtual Event, Netherlands: Association for Computing Machinery, 2021, pp. 90–94. ISBN: 9781450391283. DOI: 10.1145/3487351.3488324. URL: https://doi.org/10.1145/3487351.3488324.

[70] Joseph Henrich, Steven J Heine, and Ara Norenzayan. “Beyond WEIRD: Towards a broad-based behavioral science”. In: *Behavioral and Brain Sciences* 33.2-3 (2010), p. 111.

[71] Joseph Henrich, Steven J Heine, and Ara Norenzayan. “Most people are not WEIRD”. In: *Nature* 466.7302 (2010), pp. 29–29.

[72] Daniel Hickey et al. “Auditing elon musk’s impact on hate speech and bots”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 17. 2023, pp. 1133–1137.

[73] Stephanie Hill. “‘Definitely not in the business of wanting to be associated’: Examining public relations in a deplatformization controversy”. In: *Convergence* (2023), p. 13548565231203981.

[74] Gabriel Hine et al. “Kek, cucks, and god emperor trump: A measurement study of 4chan’s politically incorrect forum and its effects on the web”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. 1. 2017, pp. 92–101.

[75] Tin Kam Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE. 1995, pp. 278–282.

[76] Sameera Horawalavithana et al. “Mentions of Security Vulnerabilities on Reddit, Twitter and GitHub”. In: *IEEE/WIC/ACM International Conference on Web Intelligence*. ACM, Oct. 2019, pp. 200–207. ISBN: 978-1-4503-6934-3. DOI: 10.1145/3350546.3352519. URL: https://dl.acm.org/doi/10.1145/3350546.3352519.

[77] Manoel Horta Ribeiro et al. “Deplatforming did not decrease Parler users’ activity on fringe social media”. In: *PNAS nexus* 2.3 (2023), pgad035.

[78] Manoel Horta Ribeiro et al. “Do Platform Migrations Compromise Content Moderation? Evidence from r/The_Donald and r/Incels”. In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW2 (Oct. 2021). DOI: 10.1145/3476057. URL: https://doi.org/10.1145/3476057.

[79] Shagun Jhaver et al. “Online harassment and content moderation: The case of blocklists”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 25.2 (2018), pp. 1–33.

[80] Jialun Aaron Jiang et al. “A Trade-off-Centered Framework of Content Moderation”. In: *ACM Trans. Comput.-Hum. Interact.* 30.1 (Mar. 2023). ISSN: 1073-0516. DOI: 10.1145/3534929. URL: https://doi.org/10.1145/3534929.

[81] Kenneth Joseph et al. “(Mis) alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 312–324.

[82] Aditya Joshi, Pushpak Bhattacharyya, and Mark Carman. "Political issue extraction model: A novel hierarchical topic model that uses tweets by political and non-political authors". In: *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. 2016, pp. 82–90.

[83] Volker Kaibel. "On the expansion of graphs of 0/1-polytopes". In: *The Sharpest Cut: The Impact of Manfred Padberg and His Work*. SIAM, 2004, pp. 199–216.

[84] Eirini Kalliamvakou et al. "An in-depth study of the promises and perils of mining GitHub". In: *Empirical Software Engineering* 21.5 (2016), pp. 2035–2071.

[85] Eirini Kalliamvakou et al. "Open source-style collaborative development practices in commercial projects using GitHub". In: *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. Vol. 1. IEEE. 2015, pp. 574–585.

[86] Eirini Kalliamvakou et al. "The promises and perils of mining GitHub". In: *Proceedings of the 11th working conference on mining software repositories*. 2014, pp. 92–101.

[87] NG Kin Wai, Sameera Horawalavithana, and Adriana Iamnitchi. *Multi-platform information operations: Twitter, facebook and youtube against the white helmets*. 2021.

[88] Michael Klug and James P Bagrow. "Understanding the group dynamics and success of teams". In: *Royal Society Open Science* 3.4 (2016), p. 160007.

[89] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[90] Robert E Kraut and Paul Resnick. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

[91] Rohan Kshirsagar et al. "Predictive embeddings for hate speech detection on Twitter". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, 2018, pp. 26–32.

[92] Dilek Küçük and Fazli Can. "Stance detection: A survey". In: *ACM Computing Surveys (CSUR)* 53.1 (2020), pp. 1–37.

[93] Kai Kupferschmidt. "As Musk reshapes Twitter, academics ponder taking flight". In: *Science (New York, NY)* 378.6620 (2022), pp. 583–584.

[94] Lucio La Cava, Sergio Greco, and Andrea Tagarelli. "Understanding the growth of the Fediverse through the lens of Mastodon". In: *arXiv:2106.15473 [physics]* (June 29, 2021). arXiv: 2106.15473. URL: http://arxiv.org/abs/2106.15473 (visited on 08/13/2021).

[95] Lucio La Cava and Andrea Tagarelli. "Information Consumption and Boundary Spanning in Decentralized Online Social Networks: the case of Mastodon Users". In: *arXiv:2203.15752 [physics]* (Mar. 29, 2022). arXiv: 2203.15752. URL: http://arxiv.org/abs/2203.15752 (visited on 04/01/2022).

[96] Karim R Lakhani and Robert G Wolf. "Why hackers do what they do: Understanding motivation and effort in free/open source software projects". In: *Perspectives on Free and Open Source Software*. Ed. by Scott Hissam Joe Feller Brian Fitzgerald and Karim R. Lakhani. Cambridge: MIT Press, 2005.

[97] Joseph Lawrance, Seikyung Jung, and Charles Wiseman. "Git on the Cloud in the Classroom". In: *Proceeding of the 44th ACM Technical Symposium on Computer Science Education*. SIGCSE '13. New York, NY, USA: Association for Computing Machinery, 2013, pp. 639–644.

[98] John Lawrence and Chris Reed. "Argument mining: A survey". In: *Computational Linguistics* 45.4 (2020), pp. 765–818.

[99] Younghun Lee, Seunghyun Yoon, and Kyomin Jung. "Comparative Studies of Detecting Abusive Language on Twitter". In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 101–106.

[100] Josh Lerner and Jean Tirole. "Some simple economics of open source". In: *The Journal of Industrial Economics* 50.2 (2002), pp. 197–234.

[101] Antonio Lima, Luca Rossi, and Mirco Musolesi. "Coding together at scale: GitHub as a collaborative social network". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 8. 2014.

[102] Jianhua Lin. "Divergence measures based on the Shannon entropy". In: *IEEE Transactions on Information theory* 37.1 (1991), pp. 145–151.

[103] Wei-Hao Lin et al. "Which side are you on? Identifying perspectives at the document and sentence levels". In: *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. 2006, pp. 109–116.

[104] Seán Looney. "Content moderation through removal of service: Content delivery networks and extremist websites". In: *Policy & Internet* 15.4 (2023), pp. 544–558.

[105] Yuxing Ma et al. "World of code: an infrastructure for mining the universe of open source VCS data". In: *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE. 2019, pp. 143–154.

[106] Shervin Malmasi and Marcos Zampieri. "Challenges in discriminating profanity from hate speech". In: *J. Exp. Theor. Artif. Intell.* 30.2 (2018), pp. 187–202.

[107] John Matherly. “Complete guide to Shodan”. In: *Shodan, LLC (2016-02-25)* 1 (2015).

[108] Brian W Matthews. “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[109] Tobias Mayer et al. “Argument Mining on Clinical Trials.” In: *COMMA*. 2018, pp. 137–148.

[110] Amin Mekacher and Antonis Papasavva. ““I Can’t Keep It Up." A Dataset from the Defunct Voat. co News Aggregator”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 1302–1311.

[111] John Meluso et al. “Invisible Labor in Open Source Software Ecosystems”. In: *arXiv preprint arXiv:2401.06889* (2024).

[112] Ines Mergel. “Open collaboration in the public sector: The case of social coding on GitHub”. In: *Government Information Quarterly* 32.4 (2015), pp. 464–472.

[113] Bojan Mohar. “Isoperimetric numbers of graphs”. In: *Journal of Combinatorial Theory, Series B* 47.3 (Dec. 1, 1989), pp. 274–291. ISSN: 0095-8956. DOI: 10.1016/0095-8956(89)90029-4. URL: https://www.sciencedirect.com/science/article/pii/0095895689900294 (visited on 04/28/2023).

[114] Bjarke Mønsted and Sune Lehmann. “Characterizing polarization in online vaccine discourse—A large-scale study”. In: *PloS one* 17.2 (2022), e0263746.

[115] Goran Murić et al. “Collaboration Drives Individual Productivity”. In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW (Nov. 2019). DOI: 10.1145/3359176.

[116] Shawn N Murphy et al. “Current state of information technologies for the clinical research enterprise across academic medical centers”. In: *Clinical and translational science* 5.3 (2012), pp. 281–284.

[117] Matthieu Nadini et al. “Mapping the NFT revolution: market trends, trade networks, and visual features”. In: *Scientific Reports* 11.1 (Oct. 22, 2021), p. 20902. ISSN: 2045-2322. DOI: 10.1038/s41598-021-00053-8. URL: https://doi.org/10.1038/s41598-021-00053-8.

[118] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. “Random graphs with arbitrary degree distributions and their applications”. In: *Physical review E* 64.2 (2001), p. 026118.

[119] Lynnette Hui Xian Ng, Iain J Cruickshank, and Kathleen M Carley. “Coordinating Narratives Framework for cross-platform analysis in the 2021 US Capitol riots”. In: *Computational and Mathematical Organization Theory* 29.3 (2023), pp. 470–486.

[120] Abby Ohlheiser. “Fearing yet another witch hunt, Reddit bans ‘Pizzagate’”. In: *The Washington Post* (Nov. 24, 2016). URL: https://www.washingtonpost.com/news/the-intersect/wp/2016/11/23/fearing-yet-another-witch-hunt-reddit-bans-pizzagate/?noredirect=on (visited on 03/20/2021).

[121] Elinor Ostrom. “Background on the Institutional Analysis and Development Framework: Ostrom: Institutional Analysis and Development Framework”. In: *Policy Studies Journal* 39.1 (Feb. 2011), pp. 7–27. ISSN: 0190292X. DOI: 10.1111/j.1541-0072.2010.00394.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1541-0072.2010.00394.x (visited on 02/05/2022).

[122] Olga Papadopoulou, Evangelia Kartsounidou, and Symeon Papadopoulos. “COVID-related misinformation migration to BitChute and Odysee”. In: *Future Internet* 14.12 (2022), p. 350.

[123] Ji Ho Park and Pascale Fung. “One-step and Two-step Classification for Abusive Language Detection on Twitter”. In: *Proceedings of the First Workshop on Abusive Language Online*. 2017, pp. 41–45.

[124] Alexandria Payne and Vandana Singh. “Open source software use in libraries”. In: *Library Review* 59.9 (2010), pp. 708–717.

[125] Joshua M Pearce. “Building research equipment with free, open-source hardware”. In: *Science* 337.6100 (2012), pp. 1303–1304.

[126] Jeffrey Perkel. “Democratic databases: science on GitHub”. In: *Nature News* 538.7623 (2016), p. 127.

[127] Nathaniel Persily. “The 2016 US Election: Can democracy survive the internet?” In: *Journal of democracy* 28.2 (2017), pp. 63–76.

[128] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. “Effective hate-speech detection in Twitter data using recurrent neural networks”. In: *Appl. Intell.* 48.12 (2018), pp. 4730–4742.

[129] Lorenzo Prandi and Giuseppe Primiero. “Effects of misinformation diffusion during a pandemic”. In: *Applied Network Science* 5.1 (Dec. 2020), p. 82. ISSN: 2364-8228. DOI: 10.1007/s41109-020-00327-6.

[130] Ben Arfa Rabai et al. “Programming Language Use in US Academia and Industry.” In: *Informatics in Education* 14.2 (2015), pp. 143–160.

[131] EC Rajendra-Nicolucci and Ethan Zuckerman. *An Illustrated Field Guide to Social Media (p. 119). Knight First Amendment Institute*. 2021.

[132] Aravindh Raman et al. “Challenges in the Decentralised Web: The Mastodon Case”. In: *Proceedings of the Internet Measurement Conference*. IMC ’19: ACM Internet Measurement Conference. Amsterdam Netherlands: ACM, Oct. 21, 2019, pp. 217–229. ISBN: 978-1-4503-6948-0. DOI: 10.1145/3355369.3355572. URL: https://dl.acm.org/doi/10.1145/3355369.3355572 (visited on 08/13/2021).

[133] Eric Raymond. “The cathedral and the bazaar”. In: *Knowledge, Technology & Policy* 12.3 (1999), pp. 23–49.

[134] Reddit. *Homepage, Reddit INC*. 2021. URL: https://www.redditinc.com (visited on 04/04/2021).

[135] Nils Reimers et al. “Classification and clustering of arguments with contextualized word embeddings”. In: *arXiv preprint arXiv:1906.09821* (2019).

[136] Manoel Horta Ribeiro et al. “Auditing radicalization pathways on YouTube”. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 131–141.

[137] Manoel Horta Ribeiro et al. “Does Platform Migration Compromise Content Moderation? Evidence from r/The_Donald and r/Incels”. In: *arXiv preprint arXiv:2010.10397* (2020).

[138] Shahron Williams van Rooij. “Adopting open-source software applications in US higher education: A cross-disciplinary review of the literature”. In: *Review of Educational Research* 79.2 (2009), pp. 682–701.

[139] Giuseppe Russo et al. *Understanding Online Migration Decisions Following the Banning of Radical Communities*. Dec. 9, 2022. arXiv: 2212.04765[physics, stat]. URL: http://arxiv.org/abs/2212.04765 (visited on 05/09/2023).

[140] Haji Mohammad Saleem and Derek Ruths. “The aftermath of disbanding an online hateful community”. In: *arXiv preprint arXiv:1804.07354* (2018).

[141] Haji Mohammad Saleem et al. “A web of hate: Tackling hateful speech in online social spaces”. In: *arXiv preprint arXiv:1709.10159* (2017).

[142] Dominik Schlechtweg et al. “A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 732–746.

[143] Nathan Schneider et al. “Modular Politics: Toward a Governance Layer for Online Communities”. In: *Proceedings of the ACM on Human-Computer Interaction* 5 (CSCW1 Apr. 13, 2021), pp. 1–26. ISSN: 2573-0142. DOI: 10.1145/3449090. URL: https://dl.acm.org/doi/10.1145/3449090 (visited on 02/05/2022).

[144] Jose Ricardo da Silva et al. “Niche vs. breadth: Calculating expertise over time through a fine-grained analysis”. In: *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*. IEEE. 2015, pp. 409–418.

[145] Vinay Singh et al. “Aggression detection on social media text using deep neural networks”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 43–50.

[146] Parinaz Sobhani. “Stance detection and analysis in social media”. PhD thesis. Universite d’Ottawa/University of Ottawa, 2017.

[147] Davide Spadini, Maurício Aniche, and Alberto Bacchelli. “PyDriller: Python framework for mining software repositories”. In: *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2018*. New York, New York, USA: ACM Press, 2018, pp. 908–911. ISBN: 9781450355735. DOI: 10.1145/3236024.3264598.

[148] Rachele Sprugnoli et al. “Creating a whatsapp dataset to study pre-teen cyberbullying”. In: *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. 2018, pp. 51–59.

[149] Marc Stevens et al. “The first collision for full SHA-1”. In: *Annual international cryptology conference*. Springer. 2017, pp. 570–596.

[150] Carolin Strobl et al. “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC Bioinformatics* 8.1 (2007), p. 25.

[151] Bartosz Taudul. *Archiwum Polskiego Usenetu*. URL: https://usenet.nereid.pl/ (visited on 05/27/2022).

[152] Yla R Tausczik and James W Pennebaker. “The psychological meaning of words: LIWC and computerized text analysis methods”. In: *Journal of language and social psychology* 29.1 (2010), pp. 24–54.

[153] The Bluesky Team. *Moderation in a Public Commons*. June 26, 2023. URL: https://blueskyweb.xyz/blog/6-23-2023-moderation-proposals (visited on 06/26/2023).

[154] David Thiel and Miles McCain. *Gabufacturing Dissent: An in-depth analysis of Gab*. 2022.

[155] Pamela Bilo Thomas et al. “Behavior Change in Response to Subreddit Bans and External Events”. In: *arXiv preprint arXiv:2101.01793* (2021).

[156] Ferdian Thung et al. “Network structure of social coding in GitHub”. In: *2013 17th European conference on software maintenance and reengineering*. IEEE. 2013, pp. 323–326.

[157] Trang Tran and Mari Ostendorf. "Characterizing the Language of Online Communities and its Relation to Community Reception". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 1030–1035. DOI: 10.18653/v1/D16-1108. URL: https://aclanthology.org/D16-1108.

[158] Milo Trujillo et al. "What is BitChute? Characterizing the "Free Speech" alternative to YouTube". In: *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 2020, pp. 139–140.

[159] Milo Trujillo et al. "When the Echo Chamber Shatters: Examining the Use of Community-Specific Language Post-Subreddit Ban". In: *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. 2021, pp. 164–178.

[160] Milo Z Trujillo, Laurent Hébert-Dufresne, and James Bagrow. "The penumbra of open source: projects outside of centralized platforms are longer maintained, more academic and more collaborative". In: *EPJ Data Science* 11.1 (2022), p. 31.

[161] Milo Z Trujillo et al. "The MeLa BitChute Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 1342–1351.

[162] Milo Z Trujillo et al. "The MeLa BitChute Dataset". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 16. 2022, pp. 1342–1351.

[163] Adam Tutko, Austin Henley, and Audris Mockus. "More Effective Software Repository Mining". In: *arXiv preprint arXiv:2008.03439* (2020).

[164] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[165] Bertie Vidgen and Taha Yasseri. "Detecting weak and strong Islamophobic hate speech on social media". In: *J. Inf. Technol. Politics* 17.1 (2020), pp. 66–78.

[166] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[167] Felix Ming Fai Wong et al. "Quantifying political leaning from tweets, retweets, and retweeters". In: *IEEE transactions on knowledge and data engineering* 28.8 (2016), pp. 2158–2172.

[168] Lucas Wright et al. “Vectors for Counterspeech on Twitter”. In: *Proceedings of the First Workshop on Abusive Language Online*. Vancouver, BC, Canada: Association for Computational Linguistics, Aug. 2017, pp. 57–62. DOI: 10.18653/v1/W17-3009. URL: https://aclanthology.org/W17-3009.

[169] Moran Yarchi, Christian Baden, and Neta Kligler-Vilenchik. “Political Polarization on the Digital Sphere: A Cross-platform, Over-time Analysis of Interactional, Positional, and Affective Polarization on Social Media”. In: *Political Communication* 38.1–2 (Mar. 2021), pp. 98–139. ISSN: 1058-4609, 1091-7675. DOI: 10.1080/10584609.2020.1785067.

[170] Jillian C. York and Ethan Zuckerman. “Moderating the Public Sphere”. In: *Human Rights in the Age of Platforms*. The MIT Press, Nov. 2019. ISBN: 9780262353946. DOI: 10.7551/mitpress/11304.003.0012. eprint: https://direct.mit.edu/book/chapter-pdf/2259431/9780262353946_caf.pdf. URL: https://doi.org/10.7551/mitpress/11304.003.0012.

[171] Jean-Gabriel Young et al. “Which contributions count? Analysis of attribution in open source”. In: *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. IEEE. 2021, pp. 242–253.

[172] G Udny Yule. “On the methods of measuring association between two attributes”. In: *Journal of the Royal Statistical Society* 75.6 (1912), pp. 579–652.

[173] Alexey Zagalsky et al. “The emergence of Github as a collaborative platform for education”. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 2015, pp. 1906–1917.

[174] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. “PolicyKit: Building Governance in Online Communities”. In: *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST ’20: The 33rd Annual ACM Symposium on User Interface Software and Technology. Virtual Event USA: ACM, Oct. 20, 2020, pp. 365–378. ISBN: 978-1-4503-7514-6. DOI: 10.1145/3379337.3415858. URL: https://dl.acm.org/doi/10.1145/3379337.3415858 (visited on 02/05/2022).

[175] Ziqi Zhang and Lei Luo. “Hate speech detection: A solved problem? the challenging case of long tail on Twitter”. In: *Semantic Web* 10.5 (2019), pp. 925–945.

[176] Haris Bin Zia et al. “Flocking to mastodon: Tracking the great twitter migration”. In: *arXiv preprint arXiv:2302.14294* (2023).

[177] Matteo Zignani, Sabrina Gaito, and Gian Paolo Rossi. “Follow the “mastodon”: Structure and evolution of a decentralized online social network”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1. 2018, pp. 541–550.

[178] Matteo Zignani et al. “The Footprints of a “Mastodon”: How a Decentralized Architecture Influences Online Social Relationships”. In: *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). Apr. 2019, pp. 472–477. DOI: 10.1109/INFCOMW.2019.8845221.

[179] Steven Zimmerman, Udo Kruschwitz, and Chris Fox. “Improving hate speech detection with deep learning ensembles”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[180] George Kingsley Zipf. “The meaning-frequency relationship of words”. In: *The Journal of general psychology* 33.2 (1945), pp. 251–256.

[181] Nikolas Zöller, Jonathan H Morgan, and Tobias Schröder. “A topology of groups: What GitHub can tell us about online collaboration”. In: *Technological Forecasting and Social Change* 161 (2020), p. 120291.

[182] Arkaitz Zubiaga et al. “Analysing how people orient to and spread rumours in social media by looking at conversational threads”. In: *PloS one* 11.3 (2016), e0150989.

[183] Arkaitz Zubiaga et al. “Discourse-aware rumour stance classification in social media using sequential classifiers”. In: *Information Processing & Management* 54.2 (2018), pp. 273–290.

[184] Ethan Zuckerman. *A social network taxonomy*. 2023. URL: https://newpublic.substack.com/p/a-social-network-taxonomy (visited on 02/19/2023).