NORMIES AND ANONS: A STUDY OF IMPOLITENESS, SWEARWORDS, AND SENTIMENTS IN ANONYMOUS SPEECH ONLINE

_____

A Thesis

Presented to the

Faculty of

San Diego State University

_____

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Linguistics

_____

by

Keith Edward Diedrick

Spring 2016

SAN DIEGO STATE UNIVERSITY

The Undersigned Faculty Committee Approves the

Thesis of Keith Edward Diedrick:

Normies and Anons: A Study of Impoliteness, Swearwords, and Sentiments in Anonymous Speech Online

Rob Malouf, Chair
Department of Linguistics and Asian/Middle Eastern Languages

Eniko Csomay
Department of Linguistics and Asian/Middle Eastern Languages

Radmila Prislin
Department of Psychology

4/29/2016
Approval Date

ABSTRACT OF THE THESIS

Normies and Anons: A Study of Impoliteness, Swearwords, and Sentiments in Anonymous Speech Online

by
Keith Edward Diedrick
Master of Arts in Linguistics
San Diego State University, 2016

This paper examined the online phenomenon of complete anonymity during discourse in terms of how it affected (im)politeness strategies, the frequency and manner of swearword usage, and overall sentiment within the community. The overall goal of this examination was to attempt to explain whether or not anonymity caused users to become more vulgar, impolite, insulting, or overall more negative in their speech online. Many sites with communicative goals in mind are designed specifically to discourage or disallow anonymous participation specifically due to a belief that lacking identification causes users to freely insult or offend other users. Specifically, the belief is that anonymity generates a gap and breaks the connection between interlocutors and removes responsibility for one's actions.

Through data collected from 4chan.org and the similarly constructed but non-anonymous site reddit.com, the study examined swearword frequency and usage, insults, and overall Sentiment of comments to first identify trends between sites. These differences between sites were then examined in a more narrow scope on selected subsections of each site. By using the narrower scope, the data aided in the determination of whether or not the preconceived notions of 4chan.org's vulgar and impolite behavior online were truly caused by anonymity, or if another social pressure was involved.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

PAGE

ACKNOWLEDGEMENTS

I would like to thank all those without whose support I never would have made it this far. My friends and family have helped me through every step of the way, whether by giving me support or by merely listening to me drone on about the topics and work at hand. I would also like to express my utmost thanks to my Thesis committee for their support, ideas, and insight.

INTRODUCTION

The landscape of language use has been altered in recent decades by the advent of the internet, giving rise to new and varied situations for it including written informal language use. Groups of people and networks which otherwise would never have met can share ideas, speak freely to each other, and communicate effectively. Problems and questions unique to these new environments for communication have arisen just as readily as the environments themselves. These vary from questions of body language to context, but some are even more unique to internet and text based communication. Critics of internet communication and internet communities themselves often bring up another: How are we to handle impoliteness or rude behavior online, and especially how does the anonymity provided by certain websites exacerbate this or other communication issues online?

True and complete anonymity is fairly rare, even online, as it goes against many of the core design functions of websites with communicative goals (Bernstein et al., 2011). Anonymity does exist in a handful of places, despite this design push and the continued belief that disallowing anonymity engenders a sense of cooperativeness within the community. The total anonymity during full, speedy discourse afforded by these sites (as opposed to say, anonymous graffiti or notes) is only possible due to the new medium provided by internet communication. With not just body language or tone of voice missing from their communication, how or if this level of anonymity and supposed freedom from societal norms and expectations it provides alters participants' speech is an interesting question. More specifically, if core design functions of sites with communicative goals discard anonymity for believing it breaks down cooperation and generates offensive behavior or impolite behavior while the sites simultaneously lack many aspects of speech that can help interlocutors understand one another not just communicatively but also empathically, how can these mostly anonymous or only anonymous websites continue to function as communities without everyone constantly offending one another? This study aims to find

out, though we must first truly understand a bit about the sites we are collecting data from and their status in the greater internet community.

4CHAN

One of the most infamous - and commonly used - anonymous websites is 4chan.org. Well-known for spawning or inspiring the activist/hacktivist group "anonymous" and having received media attention (such as the widespread discussion of "Gamergate" or celebrity photo leaks and email hacks of 2014) for various actions from identifying criminals to manipulating contests for amusement, 4chan.org's reputation grows based on how the community acts outside the site itself. 4chan serves as a forum for discussing anything from Japanese animation to recent developments in science and technology. Despite the discussion focus, it directly eschews the established principle of using usernames or some other pseudonymity by providing all posters with the handle "anonymous" by default. This affords both the users and their communities as a whole a unique standing, creating a collective lack of identity. While it is possible to escape this default anonymity, 90% of users do not (Bernstein et al., 2011). By contrast, pseudonymous sites require a moniker be chosen by each user, creating a false identity in contrast to the normalized lack of personal identity found on 4chan. These usernames carry the potential to be recognized by other users, with some members becoming locally famous. The inclusion of usernames generates a site-specific reputation and personal identity for each user, a concept which 4chan dismisses. On pseudonymous sites, your reputation is tied to your username, while on 4chan your anonymity removes this unique identification, potentially freeing you from the concerns of reputation and identity and allowing for more trolling (or flaming), but according to some allows for the formation of a more cohesive, if potentially less cooperative, network (Rains, 2007; Tanis & Postmes, 2007).

The standard model for other websites is a construction which disallows or discourages the anonymity and ephemerality such as 4chan's. One example is that of Facebook, which ties a user's real name (though fake names are not impossible) and information and connects it to their posts. Many websites do not go this far, but in order to avoid what 4chan has wholeheartedly adopted, have adopted a form of pseudonymity through usernames or other features, which allows for the formation and maintenance of

reputations tied to each individual user instead. This feat is not impossible on 4chan, but it is abnormal as discussed above. Any sense of identity created is quickly (and easily) discarded. Despite this, 4chan seems to form a cohesive social network, achieving feats which have earned it the title "meme factory" of the internet (Poole, 2010). Even with the large amount of memes, images, fads, and trends emerging from the site to spread across other internet based communities, 4chan is seen by many to be one of the worst places on the internet. It is referenced in many online communities as either "the cesspool of the internet" or "the internet hate machine" (Bartlett, 2013; Shuman, 2007). The section of the site frequently seen as the worst of all boards on 4chan in this light is the "random" board /b/. /b/ is designated for general discussion without topic or focus, and in some communities is seen as even worse than 4chan as a whole. It is also frequently branded as a shock site, a website designed specifically to elicit shock, outrage, or disgust at its contents. Due to the prevalence of /b/'s shocking nature, in some media 4chan is as a whole referred to as a shock site, such as by McCoy in his article "4chan: The 'Shock Post' Site that Hosted the Private Jennifer Lawrence Photos" (2014). In order to compare the anonymous with the pseudonymous and determine just how true these descriptions are, a website comparable to 4chan in manner of use and construction is necessary.

REDDIT

It was to this end that reddit.com was chosen as the pseudonymous counterpart to 4chan. These two sites operate in a similar manner: individual subsections of the site are dedicated to specific interests, topics, or in a few cases, general discussion or specialized types of discussion (i.e. advice, support, questions, etc). There are differences, of course, even in simple things such as how the terminology is different, and users are identified differently. This second difference is mostly due to the anonymity of 4chan, with everyone being given the moniker "anonymous" and often referring to each other simply as "anon." Reddit not only allows users to recognize each other by their username and thus incorporates a "personal identity" use of a username, but also has built in the ability to link directly to a user's profile page by including the prefix /u/ (for user) before their name. Despite these differences, both serve as hubs of communication on varied topics and as centers for their own speech community. Overlap between communities likely does exist, but it is nearly

impossible to determine to what degree and how effectively users may integrate into either groups without individual surveys and identification and analysis of these users' posts specifically.

The offensive language and general negativity claimed to exist in 4chan is not displayed, or is at the very least downplayed, on its counterpart it seems. The title "internet hate machine" is instead replaced by the tagline "the front page of the internet." Both serve similar functions as places of discourse, and many of their areas of discussion have direct counterparts. Minor terminology differences exist (what is a board on 4chan is a subreddit on reddit), but the topics often correlate. "Boards" on 4chan are referenced by short designations, while "subreddits" usually carry a longer name and share a prefix. Both have their references built into the URL of the board/subreddit, i.e. the technology forums on the sites are: boards.4chan.org/g/ and reddit.com/r/technology. There is also no direct analog to /b/ on reddit, though numerous subreddits do carry out similar functions in giving users a place to discuss a wider variety of topics (usually in subreddit-specific manners) than the focused, special-interest boards. Reddit also has its own fair share of shock sites, individual boards dedicated to being as shocking and outrageous as possible, such as /r/wtf. /r/todayilearned, /r/news, /r/askreddit and more serve to help provide the general discussion aspect, despite having more focused topics and goals than a true "random" topic board like /b/ and lacking the shocking nature entirely. Each of these has specific goals with the discussion in mind, but that discussion typically only serves as a prompt for a more generalized conversation that forms in the comment chains below the main post. In this study, board, subreddit, and forum will be used interchangeably, as these minor differences have no real impact. The major factor between these two sites appears to be anonymity as compared to pseudonymity, rather than anything else in their structure.

GOALS

It is through this lens of a pseudonymous site like Reddit that this study aims to examine the patterns of speech of these two sites and determine first if 4chan truly does use more offensive, insulting, and/or negative language, and if the anonymity of the site is what allows for this situation, or if another social pressure is responsible for the permissiveness of impoliteness. In order to properly analyze this, it is important to be aware of the framework

within which this communication takes place. Previous studies of online communication have treated the entirety of the site as a social network (Bergs, 2006; Pérez-Sabater, 2012). Social networks for the purposes of our study and these previous online communication studies are defined by L. Milroy and Milroy as "a … web of ties that reaches out through a whole society, linking people to one another, however remote" (1992, p. 5). Reddit and 4chan appear to function as social networks as well, Reddit without a doubt. There is an added layer created by the design of both websites that also must be taken into consideration. That is to say, each of the subreddits and boards make up speech communities that while not identical to, share more in common with communities of practice rather than individual social networks. Where communities of practice focus around working together as they share a craft or profession, these communities focus around specific topics or interests and have their own guidelines for both posts (the top level post which is presented on the board itself) and comments (replies to the top level post). A general requirement for these is that those participating in the conversation be at least marginally involved in the activity being discussed. For photography subreddits or boards, for example, it is expected that each member of the forum be somehow invested in a photography hobby. This can vary, from the amateur to the professional and through the inclusion of periphery activities such as photograph editing, but it remains true that while not a rule of the board, a rule of the community is that those discussing the ascribed topic be knowledgeable and/or actively participating (Eckert, 2006). For this reason, I propose that rather than communities of practice, these could be more likened to "communities of interest" which fall somewhere between communities of practice and other speech communities. Each of these boards combines to make up a social network (J. Milroy & Milroy, 1985) for the individual sites, with a single user belonging to numerous communities of practice on the site. While not the precise manner of network creation in face-to-face communication offline, this direct conglomeration of communities of practice into a social network arises from the layout and construction of these websites. Reddit and 4chan both have a construction more rigid than offline communication, in the sense that belonging to the social network without belonging to at least one community of interest isn't possible, nor is it possible to belong to a community of interest and not belong to the larger social network.

In order to examine a particular aspect of these attributes of 4chan and anonymity perceived by more mainstream media and news sites, rather than broadly look at the negative perceptions of 4chan, this study will instead focus on the offensive nature of comments. 4chan is frequently cited as being offensive on purpose, due to their anonymity, and anonymity in general has been observed to promote offensive language or impoliteness (Santana, 2014; Shuman, 2007). This extends beyond the site itself, with calculated acts of trolling being linked back to 4chan. However, this behavior off the site is often treated as though it is merely an extension of the way the site itself operates, such as the rigging of Mountain Dew's "Dub the Dew" marketing campaign with absurd flavor names *Diabeetus* or *Fapple* which was not treated as an attack, merely 4chan acting as expected (Rosenfeld, 2012). In fact, some of the examples of "trolling" selected by previous admonitions of 4chan were instead in-jokes from the site itself used off site and taken seriously or misunderstood in general. 4chan will then often take responsibility for the offense taken, "for the lulz" whether it was intentional or not. We will, therefore, focus on offensive words, i.e. swearwords, insults, specifically on their uses and prevalence on 4chan and Reddit, as well as overall sentiment of the comments of the sites to determine if the anonymity of 4chan is truly what incites users to use more offensive language with one another by comparing it against a similarly constructed and used site that features pseudonymity instead of anonymity. The complete anonymity of 4chan is a relatively new concept in communication, and the perceptions of it may be be generated in fact, fallacy, or hyperbole.

Over the course of this study, it will become important to examine each of these questions in regards to the anonymity of 4chan and in comparison to the pseudonymity of Reddit. This paper will use data collected from specific boards and subreddits, chosen due to similar popularity and topic. It is, unfortunately, slightly limited in scope given the small cross section of the internet being examined. 4chan boasts over 7 million users (Poole, 2010), and Reddit over 250 million according to its "about" page. These are not small websites, but are by no means definitive measurements of internet communication. 4chan is a large site which uses anonymity as a core concept, but it is by no means the only one. It is, however, unique in the reputation it has garnered and the fact that posts are both anonymous and ephemeral (Bernstein et al., 2011).

As mentioned above, each site uses unique terminology. For the purposes of this paper, board/subreddit/forum may be used interchangeably, and to refer to the content of these sites a hierarchy will be referenced. A “submission” is an image, text, link, or other item which is attached to, linked to, or sent to the board by a user. A “post” is the text presented by the original poster (OP), and a “comment” is the text of a reply to either a submission, post, or another comment. “Users” are any who visit the sites or forums, and “posters” are those users who submit submissions, posts, or comments. “Posters” are in direct contrast to “lurkers.”

Due to the shocking nature of the sites examined, many of the criteria, words, phrases, and names discussed in this paper are likely to - and may have been designed to - offend.

RELATED WORKS

There exist two main areas of interest for us in previous work: anonymity itself and swearing. Anonymity has been studied before as a general social concept and somewhat in language, as well as studies on 4chan itself. Swearing is more heavily studied, and competing theories have arisen to attempt to explain the phenomenon properly.

THE INTERNET AND ANONYMITY

Previous studies that have looked into anonymity as a general concept as well as studies that examined online communication from a social network approach. These previous works have supplied this study with the background knowledge required in both broad concepts and in laying groundwork for studying online communication.

The anonymity on 4chan is not forced upon its users, but it is wholeheartedly embraced by the vast majority. With each anonymous person contributing to the discussion as a whole, issues of community cohesion may arise. How, precisely, can a community effectively operate without any knowledge of each other? As Rains (2007) says in his study of anonymous contributions in online group meetings, "anonymity may make one more comfortable participating in the group's discussion, [but] may also undermine perceptions of one's contributions" (p. 121). If this holds true even in a completely anonymous group such as 4chan, the study predicts that doubting, insulting, and even negative language may be much more prevalent as the community squabbles about credibility and perceptions of each other rather than discussing the topic at hand. The results Rains found in group work reflect earlier work into the anonymity that computers more easily provide, as face-to-face communication resulted in increased positive perceptions and an increased willingness to lend credibility to the speaker (Hiltz, Johnson, & Turoff, 1986). If anonymity truly makes speakers impersonal and lacking in credibility, one can expect the perceptions of 4chan to hold true in their discussions with each other, leading to a more impersonal approach to communication and increased questioning of each other. For the purposes of this study, this

is expected to translate into increased willingness to use offensive or insulting language, and an increased use of words used to express doubt.

How, then, does this anonymity compare to the use of pseudonyms? A study by Millen and Patterson (2003) found that not only does anonymity increase the use of disruptive and offensive communication online, the use of real names or pseudonyms promoted more polite conversations. Pseudonyms went further than that, however, as "name identity helped to ensure that participants were accountable for their words," with a user's real-life, face-to-face reputation being affected by online communications. Even in instances where face-to-face communication was less prevalent or non-existent, pseudonyms and real names were found to promote trust, cooperation, and more easily allowed individuals to enter into a new community or create a community all their own. For the purposes of this study, this means that Reddit is expected to have more polite discourse (less swearing or insulting) and to allow for more cohesive communities. The findings of this study were backed up by a more recent examination of "civility" of discourse in online news site forums (Santana, 2014). It was found that anonymity promoted not just trolling or flaming, but prompted much more racism, derogatory remarks, and a large increase of swearwords.

In this case, relevant information about social networks and speech communities as they appear online is important to examine. Speech online has been found to differ quite significantly from speech face-to-face, despite many of the same situational requirements applying (Pérez-Sabater, 2012). One of the major differences observed was the lack of a regularized genre of writing, with stylistic differences among comments and posts varying widely from user to user. Much of this was attributed to the use of the site still in the process of being regularized, and it's unknown so far if the irregular speech styles have regularized or if these remain dependent upon various communities, or if the variances were merely side effects of the site itself. This study also found that comments on Facebook were likely to use formalized, academic writing styles, and slightly less likely to use informal writing styles more conventionally used in chat rooms (Bergs, 2006). This informality may present itself in the difference between 4chan and Reddit, but without a more detailed analysis this study would fall short. Still, it is an important aspect to keep in mind during the discussion of the results.

Studies of Facebook and chat rooms are not the sole representations of discussions of online communities. 4chan itself has been the focus of one such study, oriented towards the analysis of the effects both anonymity and ephemerality have upon the community dynamics of the site, specifically focused on /b/ (Bernstein et al., 2011). Ephemerality was found to promote repeated discussions and posting of well-liked content as without these repetitive practices the popular topic would be lost as 4chan does not archive posts, while those forgotten and discarded topics or ideas are left to disappear in the rapidly changing waves of content. This was found to create a strong community focus on preserving the ideas the community liked. Anonymity was likewise found to promote a strong community identity, as a form of common identity. Bernstein's study did focus entirely on the board /b/, "because it is 4chan's first and most active board — it claims 30% of all 4chan traffic — but also because, in the words of its creator, it is the 'life force of the website'" (Bernstein et al., 2011). This limits the scope of the study, though perhaps not in a truly debilitating way. /b/ is a very popular board, and in particular it may better exemplify the effects of anonymity and ephemerality than any of the other boards. It is for this reason that this study assigns special attention to /b/ as well, though through different methodology which allows for the inclusion of other boards as well.

SWEARING

As this study will focus heavily on the usage of swearwords and insulting language, studies which examined these words and placed them in the context of face-to-face speech was examined for comparative purposes. swearwords were found to be prevalent, but dependent upon a speaker's background, emotional status, and personal experiences and motivations (Jay, 2009). Jay (2000) found that two thirds of speakers' usage of swearwords (or taboo words) were prompted by anger and frustration, eliciting epithets exclaiming these emotions effectively (e.g. *holy shit!*), or using them in insulting manners (*fuck off!*). The remaining usages covered various emotions, from surprise or pain, to being used in lewd jokes or for extreme joy. The fact that large selections of these words are used specifically to express anger, frustration, or be insulting, allows for their usage as a metric of whether the perceptions of 4chan hold true, but this study must incorporate other items to account for that remaining third.

Not only are swearwords used as utterances of anger and frustration and their use dependent upon the speaker, but they are often used as utterances meant specifically to offend. In a sense, swearwords may be used as a form of impoliteness strategy in a pragmatic sense, in a manner similar to politeness. Culpeper (2011) examined the use of swearwords in this manner, first defining what is meant by impoliteness. He details two main causes and two other major factors which are found in an impoliteness: first, an impoliteness strategy must conflict with the Hearer's social expectations of how the Speaker should be addressing the Hearer and the words cause or are presumed to be the cause of the offense. Second, the offense can be exacerbated by other factors like intentionality, and are very context-dependant. With this restrictions in mind, Culpeper posits that swearwords are considered impolite in all but very few, specific instances and as such this last factor (context) plays little to no role for them. He agrees that this context-spanning interpretation does not match up with theories of politeness strategies, but argues that the inherently impoliteness of the words themselves allow for a context-spanning use in at least this sense. He finds that the use of impoliteness strategies such as swearwords falls into one of three categories: "affective," "coercive," and "entertaining" impoliteness. Affective impoliteness focused on eliciting emotional response from the Hearer, typically anger, and the blame for these negative states being placed upon the Speaker. Coercive impoliteness instead was used by the Speaker to intentionally attempt to increase their power over the Hearer. Finally, Entertaining impoliteness focuses not on the Hearer or the Speaker, but on a third party who finds amusement in the impoliteness. These are not mutually exclusive, and could often overlap depending on intent and interpretation, and can each apply to swearwords.

Culpeper's take on context-spanning swearwords fails to account for observed phenomenon, such Stapleton's (2010) study that seems to indicate a correlation between male speech and swearword usage. In particular, Stapleton's data shows working class male speech contains the greatest count of *fuck*, leading her to argue that the use of swearwords is likely used in such a way as to invoke working class male speech.

Taking Stapleton's social interpretation further and building off of Culpeper's interpretations of these impoliteness strategies, Christie (2013) provides another take on swearwords and other taboo words. She argues that it is not strictly possible to assume swearwords are context-spanning, and instead that Stapleton was correct in assuming

indexicality of swearwords plays a large role in determining when swearword usage is acceptable. However, according to Christie, relevance theory also plays a large part in determining which of a wide variety of potential indexicality judgements are selected by the reader or hearer. Therefore, first a Hearer (or, more appropriately for newspapers and for websites, reader) must use their understanding of relevance to determine which indexicality judgements are to be made, then make them, and through this process determine their opinion of the swearword use. For this reason, Christie argues that each environment in which swearwords occur should be individually analyzed.

METHODOLOGY

This study collected various submissions and their subsequent comments from the top one hundred posts of each selected board or subreddit. Our collection was performed multiple times over several weeks, collecting the top 100 posts each time (in terms of recent popularity, not total popularity) and removing any duplicates. Each post and the complete tree of comments were automatically collected by a Python script and compiled into a large collection, categorized in numerous ways. It was then identified by a unique ID and coupled with features such as which site it came from, and which board or subreddit it was pulled from. Where applicable, each selection of text was also assigned a status as a submission, a post, or a comment. This automatic collection was done in complete compliance with Reddit and 4chan API's and Terms of Service.

Too large of a sample size would have made data analysis impossible for the timeframe of the project. As such, only a carefully determined selection of boards and subreddits could be examined as opposed to the hundreds available. In an attempt to counter the impact of this limited selection, boards and subreddits were selected with multiple factors in mind. First and foremost, a selected subreddit had to be frequented and subscribed to by at least 500,000 users, to ensure that a proper sample size could be collected. Second, a comparable board on 4chan needed to exist. Exceptions to this second factor were allowed, as some very popular subreddits and one very popular board (/b/) had no direct counterpart on the other site, but a collection of each was needed to properly depict the social network of the site. Finally, in order to provide an interesting point of comparison for the perceptions that Reddit carries of 4chan, the subreddit /r/4chan was selected as well. This subreddit is dedicated to the discussion of threads, boards, and users found on 4chan as well as the site 4chan itself. r/4chan focuses on /b/ in particular, making it the most likely counterpart to the general discussion board of 4chan. The final selections of boards and subreddits are depicted in Table 1, complete with relationships and topics.

Table 1. Subreddit and Board Selections and Their Counterparts and Topics

Reddit.com	4chan.org	Topic
r/askreddit	---	Community Questions
r/aww	/c/	Cute images
r/books	/lit/	Literature
r/explainlikeimfive	/adv/	Requests for short explanations
r/fitness	/fit/, /sp/	Fitness and Sports
r/food	/ck/	Food
r/funny		Funny images
r/gaming	/v/, /vg/, /vr/	Video games
r/history	/his/	History
r/jokes	---	Jokes
r/music	/mu/	Music
r/photography	/p/	Photography
r/science	/sci/	Science
r/todayilearned	/adv/,/diy/	Helpful tips and new knowledge
r/news	---	News
r/technology	/g/	Technology
r/wtf	---	Shocking concepts and images
---	/x/	Paranormal
---	/b/	Anything

In order to provide more accurate insights into the workings of these communities, subreddits and boards were grouped into seven total selections. Due to the direct desire of /b/ and the shocking subreddits to be both offensive and shocking, these were the focuses of this separation. The first two sections collected and separated data based solely on which website they came from, Reddit or 4chan. 4chan was then further separated to examine the effects of the inclusion of /b/, creating one group that encompassed all boards except /b/ and another that includes only /b/. In order to examine the impact of shock sites and the particular nature of r/4chan, these were each split from the bulk of Reddit, giving us the last three groups. For clarity's sake, these selections have been labeled thus in each subsequent table: 4chan - All, Reddit - All, 4chan - no /b/, Reddit - no shock, Reddit - shock, /b/, and r/4chan.

SWEARWORDS

For the first analysis of the data collected a selection of swearwords was chosen. The words were chosen from the ten most commonly used as detailed by Jay (2009) and were specifically selected for ease of counting while eliminating culturally or religiously specific swearwords in an attempt to gather data removed from casual mentions of religious icons or idols. The end result of this selection was five swearwords, in order of frequency of use: *fuck*, *shit*, *damn*, *ass*, and *bitch*. In addition to these five, a sixth word was chosen to examine due to its prevalence in the 4chan community: *faggot*. The word *faggot* (or *fag*) has so permeated 4chan that a common way to refer to people is via the terminology of *(adjective)fag*, e.g. *newfag* or *poorfag*,as referenced by the title of this study. Once the words to examine were selected, each group of subreddits/boards was examined for variations of each swearword and the different representations were counted. In order for a representation to be used in the count, it needed to make up 0.001% of the total word count (e.g. *reikashit* was not counted as it is not obvious what the usage is and makes up less than 0.0000001% of all words as it only occurs twice, likely in a specific thread and not in general usage over the whole site). This percentage approach was adopted because the total amount of content varied greatly between the seven groups of boards, and to eliminate the large amount of representations which only occurred once in the entire set of data. Most occurrences of single/non-repeated forms of swearing were collected from external links, quotes, or typos, and not indicative of usual speech on the selected boards. In one instance, this meant adopting not a percentage requirement, but an overall count requirement of at least a noticeable amount of occurrences in the entirety of the text, or in this specific instance, obvious use that excluded it from the previous collection of undesirables. While this may seem to have unnecessarily over collected in the instance of this specific subreddit, it proved necessary to appropriately compare with the other collections of boards. In order for representations to be selected in all other cases, each occurrence also needed to contain the swearword or a synonymous

swearword as well as account for 0.001% of the total words in the data set.[1] The word *badass* was excluded from the tabulation of *ass*, as it carried greatly different connotations from other uses (such as *jackass* or *asshole*) despite including the swearword. This decision was made in order to focus more directly on the question of anonymity affect patterns of speech in a negative, derogatory, or insulting manner as predicted by public media perceptions of 4chan. Finally, because of the lower word count in the smaller data sets of some board groups, the amount of representations chosen may be skewed towards the higher end. This issue was not attended to specifically because it is a side effect of the percentage approach, which gave much more accurate percent usage statistics. As the percentage showed much more useful data, the high representation counts were deemed acceptable until further study.

An attempt to place these swearwords in context was made, with the goal of accounting for the occurrences more easily taken as positive uses, or those presented as part of a quote. Due to the large amount of data collected and the relatively limited possibility of scope during the course of this particular study, a small sample of randomly selected contexts was examined. This cursory examination of context did not yield any conclusive results, and may be an interesting area for future research.

COLLOCATIONS

The second analysis of data was focused on using similar methods to attempt to overcome the limitations on contextual analysis that the previous word count study experienced. While still not possible to thoroughly examine each instance of every

[1] Note: "Cunt," "whore," and "pussy" were treated as synonyms of "bitch" instead of separate swears. This was for completeness' sake, as they all share similar variations of use, but individually are used infrequently. In order to adequately capture the extent of use of this type of offensive language, these four terms were collected together. Due to the frequency of which the meanings are used in the same manner, they will be treated as tokens of a single swear for the purpose of properly analyzing attitudes. 4chan in particular seems more prone to the use of "cunt" as opposed to "bitch." the reasons behind this may be interesting, but for now they have been combined together to give even representation to a handful of swears used more frequently by some groups than others to mean the same things, i.e. derogatory references to women and implying cowardice or weakness. Some were used to refer to body parts (*cunt, pussy*) but these usages occurred so rare as to be negligible.

swearword, another Python script was built and used to identify the probability of a given comment being or containing an insult (Sharma, 2013). This is more specific than simple word counts, as it looks for collocations which indicate insulting or derogatory language. A collection of words and their equivalents are used as standards and assigned a status of "insulting." The frequency of these words within a given string is then examined, as well as the collocations of these words. The total value of "insulting" words is weighted by these collocations and then averaged against the length of the comment. The output is a percentage of likelihood that the given comment is an insult. Unfortunately this program does not extend to how harsh the insult is, merely how likely the comment is to be an insult. The type of collocation, as well as other words present in the text, increases the probability that the given comment is an insult or not. As such, each comment was passed through in its entirety. The decision was made to follow this method as opposed to the lengthy process of looking at each individual sentence, due to the limitations of the latter approach. By looking at the entire comment, quotes, sarcasm, and similarly non-insulting uses of what may otherwise indicate an insulting statement, will not be predicted to be insults as readily as actually insulting comments. The insult detection code was run on each of the seven selections of subreddits and boards, and it was determined what percentage of each had a 50%, 75%, and 90% probability of being an insulting comment.

Building off of the method of the insult location code, another code was built to determine the most common collocations of each swearword. These collocations were automatically counted, and the top 30 most common for each grouping of forums was selected and analyzed. Unique collocations which made it into the top 30 were paid special attention to, and any truly unique collocations were added to data tables for comparison, as well as the top five collocations after going through verification of their status as truly viable collocations (i.e. eliminating those collocations which occurred only due to the removal of punctuation as opposed to intended usage by redditors or anons). Despite the sheer number of comments to look through, a sample of each collocation was examined to ensure proper retrieval and count, as well as analyze the comment the collocation came from. Expected collocations weren't excluded, for the sake of completeness and accuracy. Each of these examples examined was used to collect relevant information and statistics on each of the

subreddit groupings, if applicable. Any insight gained was then examined and percentages of collocations were paired against the total usages for each swearword found.

These collocations and their percent usage was then further compared against total word count and usage in order to shine more light upon the statistics gathered in the first part of the study. This information was used to further examine the data and findings granted by sheer word count and the comparison between the subreddits. Particularly of interest was how often are swears used to be truly offensive, insulting, or derogatory when the user was anonymous as opposed to when pseudonymous. The statistics gathered from collocations used in insulting manners was also compared to the findings of the insult-detection code and the results it gathered for each subreddit group.

SENTIMENT ANALYSIS

In order to provide a more detailed analysis of the data provided, the offensive nature of swearwords examined as detailed above was coupled with an examination of the overall nature and attitude of the various comments and replies. Similar in nature to the analysis of insults and their detection, this approach to the data focuses on using Python for Sentiment Analysis. This approach focuses on the examination of the "sentiment, or overall opinion towards the subject matter - for example, whether a product review is positive or negative" (Pang, Lee, & Vaithyanathan, 2002). Individual sentiments may not be particularly interesting for our purposes in this study, though a collective examination will likely elicit some interesting results.

In order to gather the collective data on sentiments, the Vader Sentiment Analysis source code was modified and used to analyze the collected 4chan and Reddit comments (Hutto & Gilbert, 2014). Much more sophisticated but working on a similar premise as the insult detector, the Vader Sentiment Analysis program is designed to assign words inherent sentiment values weighted individually as to how positive or negative a given word is. This sentiment analyzer parses a single sentence at a time, and for each sentence assigns weighted positivity and negativity scores to each word. Weighted values are increased for certain collocations (e.g. *fuck* in *fuck you* is weighted more negatively than in *holy fuck*). Finally, these weighted values are averaged and compared providing negativity, positivity, and neutral scores for the given sentence. For a useful but not prohibitively large analysis, a

randomized selection of five thousand comments was used for each of the previously determined group selections of boards. The random samples were then separated into appropriate sentence chunks, with each sentence analyzed. While individual comment sentiment could have been examined as a whole, this approach was not chosen due to redundancy.

Instead, each comment was split into its constituent sentences and each sentence was analyzed. The results gathered from this were then averaged for each of the subreddit groups: Reddit - All, 4chan - All, Reddit - no shock, Reddit - shock, 4chan - no /b/, /b/, and r/4chan. The averages were limited to Positive, Negative, and Neutral. On top of this Average Sentiment, a percentage of sentences which were Positive, Negative, Completely Neutral, and Ambiguous was gathered to create a better comparative analysis of the sites. Ambiguous here is those sentences which were mostly neutral but carried noticeable negativity or positivity. As many sentences carry no sentiment or, more accurately, the sentiment they carry is neutral instead of easily polarized into positive or negative, much of the results will focus on positive and negative. This is also why it was deemed necessary to separate ambiguous, in order to show where there more exist more “grey area” between positive/negative, and neutral on the various forum selections.

RESULTS

Given the three variants of data analysis performed, the results will be separated by analysis and each selection individually presented and discussed in the following section.

SWEARWORDS

4chan users swear considerably more often than previous research on the English language indicates. According to McEnery (2006) in his recent corpora study of British English, swearwords account for roughly 0.3-0.5% of speaker's words. Another study by Mehl and Pennebaker (2003) focusing on student speech patterns found a 0.5% rate of swearword usage, consistent between speakers. Using this baseline, 4chan users swear just under twice as often with 0.9% of all words being swearwords (Table 2). In comparison, Reddit users swear less often than Mehl and Pennebaker observed students swearing and low on McEnery's range, at 0.34% of all words being swearwords. Only one swearword saw more use and more representations on Reddit than on 4chan, *damn*. With one more representation and a marginal 0.002% more usage, this is negligible compared to *fuck* or *shit*, both of which 4chan used more than twice as often as Reddit. Seemingly, anonymity does cause an increased usage of swearwords, at least from a cursory glance such as this.

Splitting away /b/, shock sites, and r/4chan from the data provides further evidence for the effect of anonymity (Table 3). Without shocking subreddits and r/4chan, Reddit (under the title of Reddit - no shock) lowers its percentage of words which are swears to 0.27%**,** just barely under the minimum found by McEnery (2006). 4chan (as 4chan - no /b/) lowers its rate of usage as well, but only by a miniscule amount down to 0.88%. Even more interesting, without shocking subreddits Reddit's usage of the word *faggot* drops from the already-negligible 0.005% to complete non-occurrence. 4chan's usage remains comparatively strong without /b/, at 0.05% of all words used. While these percentages may seem small, they remain significant ($p < .001$ using a chi-squared test), to the extent that slight

Table 2. Total Percentages of Swearwords

Word	4chan – All	Reddit – All
Fuck	8 representations 0.400% of all words	8 representations 0.141% of all words
Shit	20 representations 0.340% of all words	5 representations 0.122% of all words
Damn	4 representations 0.031% of all words	5 representations 0.033% of all words
Ass	5 representations 0.041% of all words	4 representations 0.031% of all words
Bitch	6 representations 0.040% of all words	4 representations 0.011% of all words
Faggot	10 representations 0.062% of all words	2 representations 0.005% of all words
Total	53 representations 0.900% of all words	28 representations 0.340% of all words
Difference significant at p < .0001		

variations of this total word count percentage in the range of 0.01% difference translates to 5% to 10% change in actual word frequency. For comparison's sake, one of the most commonly used words in modern American English according to the Corpus of Contemporary American English, *people*, averages 0.315% of the total word count on these boards. To provide further evidence, if a perfect distribution is assumed, on average the count of words between Reddit and 4chan for these swearwords differs very significantly. 4chan's word counts of *fuck*, *shit*, and *faggot* show a 200% to 350% increase in size from an adjusted Reddit word count. By comparison, other common words such as *people*, *ever*, and *time* show 10% loss of use to 20% increase of use. Tables 2 and 3 have given us reason to believe that the perceptions of 4chan may be grounded in reality, at least in regards to the frequency of swearing. The anonymity of the website may indeed be causing this drastic increase of swearword usage. The final points of reference are to examine the data from those boards which were removed from the previous selections, /b/, r/4chan, and the shock-based subreddits. The data presented in Table 4 seem to counter the points revealed in Tables 2 and 3. /b/ has much higher rates of swearword usage than previously seen as 1.86% of all words are swearwords, more than three times the highest rates observed in both McEnery's corpora study and Mehl and Pennebaker's speech study. However, the non-anonymous subreddit dedicated to 4chan, r/4chan, bears a similarly high statistic with 1.4% of all words being swears. Even the shock subreddits come in high with 0.8% of words being some form of swearword. This is higher than the rate of 4chan without /b/, but by comparatively little. If the high rates seen in 4chan were truly caused by anonymity and nothing else, even shock subreddits and the 4chan-dedicated subreddit r/4chan would be expected to be lower than 4chan with or without /b/. Something else seems to be happening here than just anonymity.

Table 3. Percentages Excluding Shock Sites

Word	4chan – no /b/	Reddit – no shock
Fuck	9 representations 0.391% of all words	8 representations 0.111% of all words
Shit	20 representations 0.342% of all words	5 representations 0.102% of all words
Damn	4 representations 0.030% of all words	5 representations 0.031% of all words
Ass	5 representations 0.042% of all words	4 representations 0.028% of all words
Bitch	6 representations 0.032% of all words	4 representations 0.010% of all words
Faggot	9 representations 0.053% of all words	0 representations 0.000% of all words
Total	53 representations 0.882% of all words	26 representations 0.273% of all words
Difference significant at p < .001		

Table 4. Shock site, /b/, and r/4chan Percentages

Word	/b/	Reddit – shock	r/4chan
Fuck	17 representations 0.791% of all words	11 representations 0.350% of all words	12 representations 0.633% of all words
Shit	12 representations 0.364% of all words	11 representations 0.233% of all words	19 representations 0.428% of all words
Damn	6 representations 0.066% of all words	7 representations 0.057% of all words	5 representations 0.049% of all words
Ass	7 representations 0.141% of all words	6 representations 0.083% of all words	5 representations 0.060% of all words
Bitch	11 representations 0.160% of all words	6 representations 0.038% of all words	6 representations 0.069% of all words
Faggot	24 representations 0.338% of all words	7 representations 0.050% of all words	15 representations 0.170% of all words
Total	77 representations 1.861% of all words	50 representations 0.804% of all words	62 representations 1.400% of all words
Difference significant at p < .001			

As observed above, without shock subreddits or r/4chan, reddit's usage of *faggot* didn't just decrease but entirely disappeared. Perhaps this has an impact, as shown in that the removed subreddits and all 4chan selections have a high representation count for this word. A possible explanation of this is the jargon words 4chan uses as discussed previously. *Newfag* and *poorfag* are both counted in the representation for *faggot* for 4chan. If these

jargon words were removed[2] from the data, a better insight may be obtained. The new percentages for *faggot* and totals may be seen in Table 5. Readers may note that the percentage of use of shock subreddits has also changed. This is because during the examination of these jargon words, one which was previously not known to the researcher was observed: *faget*. Originally taken as a misspelling of *faggot*, it appears to instead be a subreddit-specific word occurring only on r/spacedicks. *Faget* is defined online by Urban Dictionary as an individual "engrossed in the process of faggotry. The alternate spelling is to distinguish those instigators of faggotry from legitimate homosexuals, towards whom no offense is intended" (Faget, 2010).

Without these jargon words, as seen in Table 5, very little changes in the totals. The largest total decrease is in /b/, despite Reddit - shock losing nearly 80% of all *faggot* counts. The general trend also remains relatively the same, as each subreddit loses only a small amount compared to the grand totals. It is, however, interesting to note that these jargon words appear to focus on the *faggot* category, providing us with another worthwhile area for further research: the usage of *faggot* in online communities, particularly those focused on shock reactions and being offensive for the sake of being offensive.

INSULTS AND COLLOCATIONS

Presented in table 6, the first aspect of this analysis is the percentage of comments predicted to be insulting. The data selected was split into three groups for our purposes: above 50% to 74% chance, 75% to 89% chance, and 90% chance and above that any given comment was insulting. The subreddits and boards in this table are sorted in the same manner as the previous study, with the full collection given first followed by the sum minus the most shock-oriented boards and completed by the data from those separated forums.

---

[2] And perhaps they should be, as these words are not necessarily used as swearwords, but more accurately as substitutes for "new person" or "poor person," seeking to not cause offense. In fact, variations on these words are often used to self-identify. One may call oneself *newfag* or *sportsfag* to identify one's status within the group and conversation without any negative implication.

Table 5. Adjusted *Faggot* Totals for All Selections

Board Selection	New Faggot Totals	New Cumulative Totals
4chan – All	5 representations 0.052% of all words	48 representations 0.895% of all words
Reddit – All	1 representations 0.003% of all words	27 representations 0.340% of all words
4chan – no /b/	5 representations 0.047% of all words	49 representations 0.877% of all words
Reddit – no shock	0 representations 0.000% of all words	Unchanged
/b/	8 representations 0.257% of all words	61 representations 1.78% of all words
Reddit - shock	2 representations 0.010% of all words	43 representations 0.766% of all words
r/4chan	6 representations 0.150% of all words	53 representations 1.39% of all words

4chan outperforms Reddit in all sections, almost doubling the percentage of comments which fall under each range. In line with previously discussed media assumptions about the site, 0.71% of all texts collected from 4chan have a 90% or higher likelihood of being insulting. This contrasts with Reddit, which carries half the probability at 0.35% of texts with a 90% chance of being an insult. The same comparison can be made in the lower ranges of probable insulting nature, with both sites experiencing a spike in the 50% to 74% range. This lower range is where we expect quotes of insults, insults presented in narrative or other similar discourse formats, and discussions of insults to fall. The contrast is even greater

Table 6. Percentage of All Insulting Comments in Given Ranges of Probability

Range of Probability of Insult:	**4chan – All**	**Reddit – All**	**4chan – no /b/**	**Reddit – no shock**	**/b/**	**Reddit – shock**	**r/4chan**
50% - 74%	1.52%	0.89%	1.49%	0.72%	2.18%	1.36%	2.43%
75% - 89%	0.74%	0.37%	0.72%	0.26%	1.39%	0.62%	1.45%
90% +	0.71%	0.35%	0.68%	0.20%	1.31%	0.73%	1.76%

Difference between sites marginally significant at $p < .10$

once /b/ and shocking subreddits are removed from the dataset. Without shock-focused subreddits or r/4chan, the rate of insults under Reddit - No Shock drops to 0.2% with a greater than 90% chance of being an insult. 4chan undergoes no such transformation, with only 0.03% fewer insulting texts. This difference in the decrease of insulting texts and

comments holds across each of the percentage of surety ranges. As a point of comparison, the last three groups of boards are particularly interesting. Not only do the shock subreddits and r/4chan lower the overall percentage of likely insults on Reddit, but individually their rates are very high. The shock subreddits boast 0.73% of their total texts as more than 90% likely to be an insult, placing them at the same range as 4chan as a whole. In the lower ranges, however, the shock subreddits fall behind and down to 1.36% in the 50-74% likely range. While still higher than Reddit as a collective, it doesn't quite match 4chan. On the other side, while 4chan remained relatively stable after the removal of /b/, /b/ itself has nearly double the percentage of insulting texts and comments in the middle and higher ranges. r/4chan outperforms all others, though, with a massive (and higher than even 4chan's 50-75% range percentage) 1.76% of all comments being more than 90% likely to be insulting.

A general trend seems to exist in this data, as well. All selections of boards have about half of all possible insults in this 50% to 74% likelihood range. The middle and high ranges each make up about a quarter of all possible insults. There is also a general trend for a very slight, but almost negligible drop off between the 75-89% and 90%+ ranges. The differences between these two ranges can be seen in Table 7. Two exceptions to the trend can be seen. First, Reddit's shock subreddits increase their percentage of insults in the highest likelihood range by a small amount. Second, r/4chan increased by a substantial amount, almost the entire percentage of Reddit as a collective unit in the same range. However, this trend is not significant across all subreddit groups ($p > .10$). The simple difference between each group remains more important, but marginally ($p < .10$), pointing towards less difference between the sites in the likelihood with which someone will insult another.

Table 7. Change in Percentage

Selection of Boards	Change from 75-89% range to 90%+ range
4chan – All	-0.03%
Reddit – All	-0.02%
4chan – no /b/	-0.04%
Reddit – no shock	-0.06%
/b/	-0.08%
Reddit – shock	+0.09%
r/4chan	+0.31%

An automated process which determines the likelihood of the given text being an insult is useful, but cannot stand against human decisions in terms of accuracy (even if the code is 80% accurate) (Sharma, 2013). In order to gain back this last bit of accuracy and attempt to shed light on more direct relations between word choices on these two websites through another

viewpoint, the collocations of the two most common swears were examined closely across each board selection. The remaining swears were also examined, but only those differences or points of data most striking and useful were focused on.

Fuck, complete with representation *fucking*, is by far the most commonly used swear on both 4chan and reddit, with *shit* coming in second, and for a handful of boards temporarily usurping *fuck*'s crown. The five most commonly used collocations are given as phrases in Table 8, arranged along with the ranking of these particular representations of collocations against all others for their given swearword. In this table, *fuck* collocates are examined before and after the word due to the existence of both *fuck* as a verb and as a noun, while *shit* collocates were selected only from those preceding it, as no useful collocation information was found in the examination of the following collocates, merely that *shit* is frequently used to end a phrase. *Fucking* was similarly separated out from *fuck* and only the collocates following it were examined. "Fucking (NP)" is a commonly used phrase, but each collocate appeared to be a variation upon it (e.g. "you fucking can't," "a fucking bird"). Though certain things are obvious here, such as the fact that "the fuck" seems to be the most common pre-fuck collocate, and even with the considerations above taken into account, it can be difficult to judge just what these points of data mean from rankings in these various orders. With this in mind, Table 4.8 assigns meaning to these collocations by grouping them together into various groups and the group percentage by boards. The groups depicted in Table 9 were chosen for the various collocations based upon the most commonly used phrases observed in the data. Seven recognizable categories were created, Disbelief, Degree, Emphatic, Dismissive, Offensive, Positive, and Ambiguous. Disbelief contains those collocates found mostly in expressions of shock, distress, surprise, or doubt ("holy fuck," "what the fuck"). Degree only contains one collocate, "as fuck," such as in the sentence "He was angry as fuck." Emphatic expressions were those which contained swearwords but had no intention to insult, deride, dismiss, but instead were used to emphasize what was being said through an attention grabbing word like a swear. "Fucking with" something or someone is merely a vulgar expression ranging from absentmindedly fiddling with something to playing mind games with someone. Dismissive contains those phrases intended to reject the influence of what is being spoken of, or to express the

Table 8. Most Common Collocations by Swearword

	Preceding Fuck Collocates	Following Fuck Collocates	Shit Collocates	Fucking Collocates
4chan – All	1. the fuck	1. fuck off	1. a shit	1. fucking shit
	2. as fuck	2. fuck you	2. this shit	2. fucking retarded
	3. to fuck	3. fuck is	3. holy shit	3. fucking stupid
	4. a fuck	4. fuck up	4. that shit	4. fucking hate
	5. holy fuck	5. fuck that	5. of shit	5. fucking up
Reddit – All	1. the fuck	1. fuck you	1. holy shit	1. fucking stupid
	2. as fuck	2. fuck up	2. that shit	2. fucking hate
	3. to fuck	3. fuck that	3. a shit	3. fucking love
	4. a fuck	4. fuck it	4. of shit	4. fucking with
	5. holy fuck	5. fuck off	5. the shit	5. fucking idiot
4chan – no /b/	1. the fuck	1. fuck off	1. a shit	1. fucking shit
	2. as fuck	2. fuck you	2. this shit	2. fucking retarded
	3. to fuck	3. fuck is	3. that shit	3. fucking stupid
	4. a fuck	4. fuck up	4. holy shit	4. fucking hate
	5. holy fuck	5. fuck out	5. of shit	5. fucking up
Reddit – no shock	1. the fuck	1. fuck you	1. holy shit	1. fucking with
	2. as fuck	2. fuck up	2. a shit	2. fucking love
	3. to fuck	3. fuck that	3. of shit	3. fucking hate
	4. a fuck	4. fuck it	4. that shit	4. fucking amazing
	5. and fuck	5. fuck off	5. the shit	5. fucking stupid
/b/	1. the fuck	1. fuck off	1. this shit	1. fucking love
	2. as fuck	2. fuck you	2. holy shit	2. fucking hot
	3. a fuck	3. fuck her	3. that shit	3. fucking faggot
	4. holy fuck	4. fuck up	4. of shit	4. fucking Christ
	5. you fuck	5. fuck that	5. a shit	5. fucking moron
Reddit – shock	1. the fuck	1. fuck you	1. holy shit	1. fucking stupid
	2. as fuck	2. fuck that	2. that shit	2. fucking a
	3. holy fuck	3. fuck is	3. this shit	3. fucking weird
	4. and fuck	4. that it	4. the shit	4. fucking awesome
	5. a fuck	5. fuck up	5. of shit	5. fucking hilarious
r/4chan	1. the fuck	1. fuck you	1. of shit	1. fucking faggot
	2. to fuck	2. fuck off	2. holy shit	2. fucking cuck
	3. a fuck	3. fuck is	3. a shit	3. fucking a
	4. as fuck	4. fuck up	4. this shit	4. fucking crop
	5. holy fuck	5. fuck does	5. that shit	5. fucking christ

Table 9. Collocations Categorized

Group Name	Collocates
Disbelief	The fuck, holy fuck, holy shit, fuck is, fuck does, fucking christ
Degree	As fuck
Emphatic	Fucking with, fucking up, fucking crop, to fuck
Dismissive	Fuck that, that shit, this shit, a fuck, a shit, fuck it
Offensive	Of shit, fuck you, fuck off, fuck out, fucking stupid, fucking hate, fucking idiot, fucking shit, fucking retarded, fucking stupid, fucking weird, fucking faggot, fucking moron, fucking cuck, you fuck, fuck her
Positive	The shit, fucking love, fucking amazing, fucking hilarious, fucking awesome, fucking hot
Ambiguous	And fuck, fuck up, fucking a

speaker's lack of interest or investment (e.g. "I don't give a shit."). The largest group of collocates was, of course, those deemed Offensive with intent to insult. The most blatant of these is "fuck you." A handful of collocates are also used to express positive opinions, such as "fucking awesome." And finally, three of the collocations were difficult to place in the previous categories and were these deemed Ambiguous. "And fuck" could be followed by a dismissive or insulting phrase, for example "and fuck you, too," compared to "and fuck that, I can't afford it." The phrase around "fuck up" can be similarly interpreted, from "Shut the fuck up" to "I had a fuck up at work: I spilled coffee on my boss." The last is "fucking a," which can be either a very positive sentiment "fucking A!" as in "fucking awesome" or can precede a phrase and lend it a vulgar emphasis. The percentages for these different groups as described here can be found in Table 10.

Due to the frequency which the collocations were used in specific manners, they were sorted under created labels representative of the most prominent use of the collocations contained within. This was done to increase readability for this study, and as such is not definitive. Nearly all collocates are, to some extent, ambiguous, but these groups were selected based on majority usage (i.e. while "the shit" may be used in a phrase such as "kick the shit out of him" as opposed to the positive use "this is the shit," the latter proved far more common in collected samples). Some collocates did not have a significant enough majority of observed uses to be properly categorized, and so were given a unique category to properly represent the widespread use of various collocates. This Ambiguous category was still tallied and analyzed like the others, and may provide interesting results on the frequency which swearwords are used with a variety of meaning.

Table 10. Percentage of Top Collocations Sorted by Category

	Disbelief	Degree	Emphatic	Dismissive	Offensive	Positive	Ambiguous
4chan – All	33.3%	9.7%	5.1%	10.9%	24.2%	0.0%	0.8%
Reddit – All	29.1%	4.9%	3.7%	23.4%	17.5%	7.2%	3.5%
4chan – no /b/	33.4%	9.7%	3.9%	21.7%	23.2%	0.0%	3.9%
Reddit– no shock	28.6%	5.6%	5.1%	28.1%	10.3%	10.0%	5.6%
/b/	27.9%	9.2%	0.0%	24.4%	32.0%	3.1%	3.3%
Reddit – shock	42.4%	4.9%	0.0%	26.1%	13.5%	7.1%	5.9%
r/4chan	38.9%	4.3%	6.6%	20.9%	24.6%	0.0%	4.7%

Difference significant at $p < .001$

The selections were made based on observed uses and their intentions, and these selections can be further combined into three groups: Positive, Negative, and Neutral. Starting with the Neutral group, we have Disbelief and Emphatic. Collocates fell under Disbelief if they were used primarily to express emotions of shock, surprise, astonishment, and of course disbelief. The use of a swearword for the mere sake of attaching potential emphasis or conveying the sense of a strong emotion attached to whatever speech around the swearword was deemed Emphatic. In a sense, the collocations that fell under here are those used for the sake of using a swearword and the attention, focus, and potentially extreme associates the given swearword brings to the discourse. The use of Emphatic collocates varied significantly between positive, negative, and neutral uses, but all seem to share in this use of the swearword for emphasis. Within the Positive selection are Degree and Positives. The Degree group only contained one collocate, "as fuck" commonly used in phrases similar to "that's cool as fuck," though other collocates that did not appear under our top selection may fall under here, such as the similar "as shit." While positive or negative sentiments can be used with the degree, the degree itself is a positive use of the swearword. For this reason it was grouped with Positive. Collocations fell under Positive if they were used not to insult or deride, but specifically to lend a positive emphasis or term to the phrase. Chief among these was "the shit," commonly used in phrases as a positively emphasizing way of saying "the best." Counter to the two Positive groups is the Negative group, consisting of Offensive and Dismissive. These were separated by their intended use, with Offensives being used

primarily with the intention to offend or insult, and Dismissive being used to negate the importance of an idea or object, or attempt to forcefully eject a person from a conversation (e.g. “GTFO” or “get the fuck out”). Finally, those collocates which showed a wider variety of use and no strong tendency to be used in any specific manner were instead placed under Ambiguous. Those which fall under Ambiguous often shared roughly even distribution between two common uses, or were used in so many different ways that it was difficult or impossible to appropriately attribute them to any other selection of collocations.

Expressions most frequently fell under Disbelief, Dismissive, or Offensive, as to be expected given these seem to be the more common uses of swearwords in general. 4chan in general swore more, as observed previously, but bore no uses of Positives in the top 10 collocates and surprisingly used Dismissives less than half as often as Reddit. 4chan did, however, use more expressions of Disbelief and Offensive collocates, gaining substantially in both categories. With shock sites once again removed, this trend between Disbelief and Offensive remains strong, though 4chan - no /b/ increases its use of Dismissives drastically, closing the gap between the two sites. Finally, the last three selections have been where trends break down in shock site comparisons. For collocates, however, it is a slightly different story. Between /b/ and the shock subreddits, the use of Dismissives has mostly evened out, but a stark difference between Disbeliefs and Offensives emerges. Reddit - shock boasts 42.4% of all studied collocates as Disbelief, and /b/ falls behind at 27.9%. /b/ is on track with the usage across other parts of 4chan, but the shock subreddits’ use of Disbelief swear collocates has spiked. Also of note is that use of Offensives while higher than the rest of 4chan, Reddit - shock has remained in a roughly similar area to the rest of Reddit, remaining below the rate observed in Reddit as a whole. The last selection of boards, r/4chan, instead sees fewer uses of Disbeliefs, but still roughly similar to other areas of Reddit, and a large increase in the number of Offensives, pushing its percentage distribution closer to 4chan than Reddit. A further examination of these collocates and their uses should likely be conducted, analyzing each occurrence and its intended use in the discourse.

Sentiment Analysis

For further information of the overall landscape of each site and each subsection, Sentiment Analysis was, as discussed above, the next step. Each examination of swearwords

thus far has pointed towards the last group, the shock sites, /b/, and r/4chan, as the area where existing trends and patterns between our anonymous and pseudonymous users break down. Our average of the sentiments found on these sites follows a similar trend, as seen in Table 11.

Table 11. Average Sentiment Analysis Scores

	Average Negativity	Average Positivity	Average Neutral
4chan – All	0.0839	0.1082	0.8053
Reddit – All	0.0733	0.1208	0.8016
4chan – no /b/	0.0740	0.0998	0.8247
Reddit – no shock	0.0743	0.1213	0.8000
/b/	0.0965	0.0944	0.7964
Reddit – shock	0.1053	0.1030	0.7834
r/4chan	0.1086	0.1049	0.7769
Difference significant at $p < .006$			

Average positive sentiment varied within a small range, from /b/ at the lowest with 0.0944 to Reddit - no shock at the highest with 0.1213. With a variation of 0.0269 between the two extremes, this is a fairly wide range to fall on. 4chan and its boards tend to make up the lower end of the scale, with the highest 4chan sample being 4chan - All at 0.1082. While 4chan - All is more positive on average than Reddit - shock and r/4chan, it lags behind Reddit - All and Reddit - no shock. Average negativity boasts a wider range with a variation of 0.0353, and a somewhat different tale. As a whole, only /b/, Reddit - shock, and r/4chan were more negative than positive, with r/4chan being the most negative at 0.1086 average negative sentiment. Reddit - all claimed lowest average negative sentiment at 0.0733, beating 4chan - no /b/ by 0.0007. Rather than what was seen with average positive sentiments, 4chan does not take half the range itself. Instead, negativity seems to be confined to those sites found to have more offensive reputations online. /b/, Reddit - shock, and r/4chan hold the top ranges of the negativity scale, while the lower end is filled by the total collections and the collections without respective shock sites.

Despite the small difference between 4chan - all and Reddit - all in average negativity, a larger and more significant difference exists in the percentage of all sentences with negative sentiment ($p < .001$). These percentages can be seen in Table 12. 4chan - All comes in comparatively high at 3.8% negative, while Reddit - All, 4chan - no /b/, and Reddit - no shock all have 2.8% - 2.9% negative. As a whole, 4chan appears to be noticeably more

negative in their discourse. However, it seems that this is likely due to the influence of /b/. /b/ has the second highest percentage of negative comments, 6.0%. Only r/4chan comes in higher. Also of significant note, despite the increased negativity, the shock groups /b/, Reddit - shock, and r/4chan all share similar percentages of positive comments, with /b/ actually having the highest. Only 4chan - no /b/ has a percentage of sentences with positive sentiment below 5%.

Table 12. Percentages of Comments by Sentiment Held

	Percent Negative	**Percent Positive**	**Percent Neutral**	**Percent Ambiguous**
4chan – All	3.8%	5.1%	64.8%	25.8%
Reddit – All	2.8%	5.8%	66.5%	24.3%
4chan – no /b/	2.9%	3.9%	68.6%	24.2%
Reddit – no shock	2.9%	6.1%	65.7%	24.7%
/b/	6.0%	6.5%	64%	22.1%
Reddit – shock	5.9%	5.4%	63.1%	24.5%
r/4chan	6.9%	5.9%	60.6%	25.5%
Difference significant at $p < .006$				

DISCUSSION

The data presented above generates some questions as to the perceptions of 4chan in the media. The "internet hate machine" is seen as intensely disruptive, rude or offensive, and uncivil in their discourse. These traits have all been at one time or another attributed to the fact that 4chan is a site in which anonymity is not just an option, but has taken the role of the default identity. In order to examine the validity of these attributions, the data above looked for signs of offensive behavior through swearwords, insults, and prevailing negative attitudes.

As a general yardstick for "offensiveness," swearwords were examined in bulk to assess just how often each site swore in their comments and comments as a whole were analyzed to determine the probability of being insulting. Between the two sites Reddit and 4chan, one primarily pseudonymous the other primarily anonymous, the assumption seemed valid at first. 4chan anons definitely swore more frequently than their redditor counterparts. On average, 0.27% of all words used by redditors were in the selected representation of five of the most commonly used swearwords, and 0.35% of all comments were deemed 90% or more likely to be insults (Jay, 2009). This puts them just below the average range of face-to-face communication, well within what may be called generally "inoffensive" and "civil" communication, it would seem. Compare that with 4chan, with 0.90% of their word choice falling under our selected swearwords, nearly double the highest point on the observed face-to-face scale, and 0.71% of all comments highly likely to be insults. From just this counting method and insult detection, anonymity appears to present itself through more offensive word choice.

However, more must be taken into account to properly examine these two sites in a comparison of data than just sheer numbers. 4chan and reddit, while similar in structure, are two very different communities with different group identities, social pressures, and environments. To stop at just a count of swearwords and deem 4chan more offensive due to

its anonymity is to discredit the fact that 4chan is a community, and their group identity may play into this difference between these two sites. For this reason, a choice of reddit (and thus primarily pseudonymous) subreddits was made which included two which had similar reputations on reddit for their "offensive" content in both the primary posts and in comments: r/wtf and r/spacedicks. A subreddit dedicated to 4chan itself, r/4chan, was also included, in order to examine the community through the eyes of those not anonymous, but still members, imitators, or mockers of the original website

With the inclusion of these new "shock" subreddits, and the decision to examine the data minus those forums which carried the most reputation as being "offensive" or "vulgar," (such as /b/ on 4chan), the certainty of the basic word count approach dwindles significantly. On r/wtf and r/spacedicks, users swore nearly as much as the whole of 4chan, 0.80% of all words being swearwords. They also insulted each other more often, about even with 4chan as a whole, at 0.73% of all comments more than 90% likely to be insulting. r/4chan took it a step further, this pseudonymous forum going above the totals of 4chan and well on its way to matching /b/ in the usage of swearwords and rocketing to first with 1.76% of all comments 90% or more probable to be an insult. It was through this added examination that anonymity began to fall by the wayside in terms of influence over offensive language use. Instead, there was something about these communities which increased their rate of swearing and insulting.

These communities each carry a reputation in some way or another, either among other social or news websites, or within their own larger community. One user on reddit describes r/spacedicks as "[A] pointlessly shitty 'edgy' sub that was absolutely desperate to be /b/," likening the attitudes presented there with not just 4chan but with the general chat forum frequently seen as being worse than the rest of the site, /b/. Similarly, r/wtf has been identified as "the underbelly of weirdness on Reddit" (Klima, 2015). And so perhaps it is communities which specialize in shocking or offensive content that speak in shocking or offensive ways.

The analysis of collocations with swearwords initially seemed to support this idea. With Reddit boasting more Positive and Dismissive uses, and less Offensive uses, it seems that 4chan's usage of swearwords may be geared away from the uses of swearwords which serve to help a speaker identify with the group they are speaking with (positive and negative references to other people or things, rather than each other) and be more likely to

purposefully offend and insult. Again, though, we must avoid falling prey to oversimplifying these communities into "anonymous" and "not anonymous" without further investigation. Between the shock subreddits, /b/, and r/4chan, if the effect of anonymity were truly an increase in offensive or vulgar language, the expectation would be that these groups would be high on /b/ and low on Reddit - shock and r/4chan. Particularly, we would look to see that Reddit - shock and r/4chan remain lower than 4chan collectively. This would provide strong evidence that the offensive use of swearwords was solely the domain of anonymous users. Indeed /b/ does swear more than both the Reddit boards when it comes to Offensive terms. However, all three match pretty closely when it comes to Dismissive, and while Reddit - shock falls low on the use of Offensive word pairs, r/4chan does not. It matches almost exactly to 4chan - All. Perhaps most striking, though, is the massive use of Disbelief word pairs in Reddit - shock. This is potentially explained by how the very name of one of these two subreddits, r/wtf or "what the fuck," makes up a Disbelief collocation. That theory raises a problem, though. If usage changes emerge from the goal or topic of the forum, why should community pressures not play a part in this, too?

In fact, it seems they do. A large part of belonging to a social network is adopting the speech of that network, including words which may only be used frequently by that community (e.g. "newfag" on 4chan), in-jokes[3], and uses of common words specific to the community. The community and object of the forum r/wtf is focused on those images and stories which will elicit disbelief, horror, surprise, or shock. In a community where this is not just expected, but the intended goal, the usage takes on other meanings, and its originally intended meaning is frequently not worthy of reaction - it is, after all, expected that you will say "what the fuck" to a post on r/wtf. An example is these two instances of the phrase from r/wtf in examples 1 and 2:

1a. "This little comic made me say what the fuck as well"

1b. "No problem, gets me everytime as well"

---

[3] In fact, a subreddit exists specifically to help users assimilate into the community by learning about Reddit specific in-jokes, /r/outoftheloop.

2a. “what the fuck…”

2b. “Truly Wtf”

These instances seem to indicate that expressing disbelief or shock through swearing in a community dedicated to shocking material is not offensive, vulgar, or inappropriate. Instead, it is taken as matter of fact, something wholly normal for a member of this community to utter in response to a post.

While not surprising, if this same thought can be applied to other communities, the role of swearing itself may change from community to community. If this is the case, the categories used in our data are meaningless in examining communities in which interpretations, rules, or goals behind the use of various swearwords could differ drastically.. Even if normally offensive, a collocation may take on a different goal within a community which allows for that. Take for example, the two occurrences of “fuck” in examples 3, one from Reddit and the other from 4chan, each with the first responses. First the conversation from Reddit in 3a-b:

3a. “As the father of an autistic child...fuck you…”

3b. “As an autistic man . . . Fuck you too. You're the reason people assume we're all stuck up assholes who can't take a joke.”

And now a conversation with similar use of *fuck* on 4chan in 3c-d:

3c. /g/ I bought into the cancer and just got my first Macbook Pro. Was wondering this though, what are the first things I should do right after the first power on?

> inb4 muh linux install

> inb4 macbook is trash, i know but fuck you i wanted this.

3d. install muh ram

In the above examples, it is important to take special note of the difference in response attitude between Reddit and 4chan. What is more commonly taken to be truly offensive on Reddit and elicited numerous responses of returned aggression or defense of the original poster is instead replied to as normal on 4chan, or outright ignored. Swears may be used without intent to offend on Reddit as well as on 4chan to express any variation of emotions or simply as intensifiers, but from the data it seems that the most common on 4chan are outright offensive or insulting uses. Many of these are in calling someone “fucking stupid” or

"fucking retarded," but the top usage remains "fuck you." Casual swearing seems to be used often not to offend those involved, but to invite other members of the group to share in the experience and help to build the speaker's identity as part of the group despite being anonymous. The quote in example 4 from a poster on /lit/ is peppered with language that would be deemed offensive in many face-to-face situations, and would be an over-the-top response on r/books.

4. "God dammit I'm reading this book now and I'm really enjoying The Navidson Report but every fucking time Truant starts writing it's so unbearably try-hard and edgy that I want to skip it but I can't bring myself to skip part of the narrative. I don't want to hear about his fucking wacky paranoia anymore"

In the context of a 4chan board like /lit/, the liberal usage of swearing and the inclusion of a jargon word at least specific to shock sites if not focused on 4chan, "try-hard," helps to broadcast that the writer of this comment is a member of the 4chan community through appropriate use of slang. This is a very important aspect of establishing group identity (Eble, 1996). It is made even more important by the very anonymity that earlier perceptions from the media believed to cause the swearing on 4chan. Further evidence that this swearing is used to denote in-group status is the response received by those offended by its use. Those who out themselves as out-group through taking offense or reacting defensively to swearing or offensive language and behavior will find themselves being quickly identified as a "newfag" or "butthurt." This is exemplified by example 5 below, in which a poster gets called out for being unaware of these requirements and in response to a repeating offensive anon posts:

5a. >>7480580

can you just stop

The response is immediate:

5b. >>7480582

>newfag detected

The same responses are frequently given for posts which contain swears, but are used in ways meant to actually cause direct offense or are very defensive replies to perceived slights. Users unaware of the correct way of using these swearwords, the wide and varied slang, and unable to perform certain tasks (two most prominent are "triforcing" referring to the usage of specific unicode to properly align an image of three triangles into the "Tri-Force" from

popular game Legend of Zelda, and "greentexting" or quoting text from another comment) are called out as "newfags" and often told to "lurk moar." In short, users should refrain from joining the conversation until properly aware of what is expected for participants.

Swearing seems to be part of 4chan's group identity, something they do to forge strong bonds with each other and maintain cohesion in a community without names or permanent identifying markers. This isn't exclusive to 4chan, though the community's expectations likely drive increased usage of swearwords. The same can be said of r/wtf and r/spacedicks, the shock subreddits. On both of these forums, the community expectations and group identity call for increased use of swearwords. From "fucking fagets" on r/spacedicks to expressions of shock, disbelief, and horror on r/wtf, swearwords play a large role in the speech of these communities.

The evidence of the Sentiment Analysis also backs up this idea that not only is the offensive nature of 4chan not easily attributed to anonymity, the negativity present in many of the comments is spread to pseudonymous sites with similar conceptualizations and cultures. Without a doubt bearing more negative sentiment than 4chan without /b/, or Reddit without shock subreddits, the subreddits r/wtf, r/spacedicks, and r/4chan at least match /b/ and exceed the rest of 4chan. Most comments are still inherently neutral, but the negative sentiment does not seem attributable to anonymity in this comparison, present as it is in communities which require pseudonyms but have created cultures of shock and/or offensive behavior. Not only does anonymity not play as large a role in the use of swearwords and offensive terms as the media portrays, but it also does not create an identifiably more negative sentiment within the anonymous community. Instead, we see time and again that the shock subreddits r/wtf and r/spacedicks closely mirror the results from 4chan and even /b/ at times, while r/4chan often exceeds them all.

This unique case of r/4chan provides the last evidence for this theory, it seems. The community itself is dedicated to 4chan as a form of imitation or slightly-safer emulation of the notorious anonymous site. The rates of swearing and the percentage of negative comments seem to be caused by a form of "hypercorrection". Hypercorrection is "an incorrect usage [in this case of terms and mannerisms] that results from the over-application of a perceived rule" (Labov, 1968). Users of r/4chan are aware of the mannerisms and word choices of 4chan, as well as particular slang (such as "newfag," or "tripfag"). However,

frequently when a user manages to almost match the usage of a common meme or mannerism of 4chan but either goes overboard or misses part of the subtext they are often called out on this imitator status, as in example 6:

6a. how is it, /b/ros[4]

6b. >>674439582

you really mad an image out of this…

6c. >>674440285

how else is he gonna get that sweet reddit karma on r/4chan?

The treatment of these mimics is different from those known to be outsiders (or "normies") and those known to be uninitiated but potential members of the community ("newfags"). This hypercorrection of meme, image, and speech usage marks the users of r/4chan during their forays into 4chan (and in particular, /b/). It seems that their perception of 4chan's usage of swearwords and other offensive imagery, as well as overall negative sentiments, may be a misrepresentation generated from the stereotype. This would account for the increased rate of swearing, insults, and negative sentiment found on r/4chan. It would also lend strong support to the idea that the prevalence of these swears, insults, and negative sentiments that actually do exist on 4chan are a result of community and group pressures to maintain the cohesion and identity, rather than a direct result of anonymity. In the sense of Christie (2013), the readers of 4chan have a specific indexicality judgement for swearword use attached to them which allows for not only casual swearing, but frequent and seemingly unlimited use if used properly. As such, I argue that not only does this swearing index 4chan, but within 4chan itself provides a way for anon to recognize another real anon from any number of other site visitors like redditors, normies, or newfags.

Anonymity may have played a part in the development of the culture of 4chan, however. It just as easily may not have, as the shock subreddits, r/spacedicks in particular, developed similarly intentionally outrageous behavior in an environment which shunned

anonymity for the use of pseudonyms. These shock sites also carry a reputation in their own communities and in the larger internet community as a whole for being shocking and offensive. The desire to maintain this community reputation may override any desire users may have to preserve their own reputation from the "taint" of being vulgar, offensive, and/or rude. The desire to belong to the chosen network may exert a greater force over the individual, pushing them to adopt the mannerisms of the community.

It is not necessarily the culture itself that anonymity may affect or cause the behavior of, but it may enforce and enhance those already defined characteristics of a site like 4chan. Recent studies of group identity in anonymous communities show that such group dynamics, "mob mentality", and social network pressures often create much stronger bonds between those within them, eliminating much of the individual in terms of reputation in favor of the group as a united entity (Kraut et al., 2012). It may be this push to belong to the site that causes users to act this way. Another idea, however, follows that not only are group dynamics at play, but the anonymity of 4chan also allows users unlimited second chances and the ability to say or speak in ways which they wish to be would rather not do to preserve individual reputations (Bernstein et al., 2011; Dibbell, 2010). This second explanation goes against the findings of this study, in that pseudonymous sites seem to be just as able to be offensive or shocking, likely without care for individual reputation for the sake of maintaining the reputation of the collective. Indeed, it seems most likely that 4chan and other shock sites have adopted this manner of speaking as a way of identifying themselves to each other as true members. The infamy of 4chan has enregistered, as detailed as possible by Squires (2010), the speech to that specific site, leading many to believe that what is unique about 4chan has caused this speech. This is supported by the fact that while anonymity may have strengthened this style's development, it is by no means necessary as evidenced by the comparable styles of non-anonymous subreddits r/wtf, r/4chan, and r/spacedicks.

---

[4] Not included in this comment is the screen capture and offensive image attached to the post. The over-the-top nature of this image is what the following comments are using to call the imitator out on, coupled with the slang identifier to users on the board /b/.

CONCLUSION

4chan does use more offensive language, and does have a somewhat more negative overall sentiment. Is this caused by anonymity, or something else? Our data and understanding of the nature of various sites seems to indicate that it is not anonymity, but the community which generates this pattern. Perceptions of 4chan and anonymity online vary from the semi-positive moniker "internet meme factory" to reference its important role in many online communities, to the vile "cesspool of the internet" where many offensive conversations are shared. The focus here, though, is on anonymity as the culprit; is anonymity the cause of 4chan's offensive language and behavior? Studies exist on the periphery of this question, such as Santana's (2014) examination of anonymous responses to news articles and the "civility" of these comments, or Bernstein et al.'s study of how the combination of anonymity and ephemerality plays a role in the continued existence of /b/ as an online community but does not ask whether the offensive nature is truly due to the anonymity or not. Our study examined three areas of speech on 4chan and Reddit, comparing the anonymous site with its pseudonymous counterpart, in order to help answer the question of whether it truly is the anonymity which allows for the generation of such offensive language and behavior: Insults, swearwords, and Negative/Positive Sentiment.

For every metric examined 4chan exceeds Reddit, lending credence to the preconceived notion of 4chan's offensive behavior, as at face value it would appear that the choice between anonymity and pseudonymity is the only major factor. Our inclusion of an examination of subreddits dedicated to shocking images or behavior without the requirement of anonymity exist on Reddit. Furthermore, these shock sites match or exceed the rates 4chan in each metric. For equality's sake, /b/ was also removed, as /b/ is much more offensive in both language and behavior than the rest of 4chan. With these measures in place, it becomes evident that anonymity alone cannot be responsible for the offensive language and negative sentiment present on 4chan in greater amounts than in Reddit as a whole. To further

complicate this, one of the subreddits examined was r/4chan, a subreddit dedicated to images, stories, and discussion of 4chan (primarily /b/).

Anonymity of the users is not the most likely of reasons for the speech of 4chan. Instead, this study postulates that the cause lies within the community of the site. Rather than being offensive because they can get away with it through their anonymity, the group identity of 4chan's community involves the use of not only site-specific slang, but swearwords, insults, and negative sentiments. To add to this, it seems that the community has adopted them as a way of not just identifying as part of the group, but of reinforcing this group cohesion. Offensive language on 4chan is used less often with the actual goal of offending. Instead, finding offense in the language or images used on the site is a surefire way to be identified as not just a "newfag," but an outright "normie" and thus not a part of the group in the slightest of senses. This is backed up by our troublesome case of r/4chan, as r/4chan exceeds not just 4chan, but /b/ and other shock sites in its use of negative sentiment, insults, and swearwords. If this frequency of offensive language found on 4chan is due to the group identity, it isn't difficult to believe that the imitators and admirers of r/4chan will hypercorrect in an attempt to match their perceptions of 4chan.

It seems, therefore, that anonymity does not play a large role in this aspect of the speech of 4chan, but it is instead driven by a strongly cohesive social network. Such a network itself may be reinforced or promoted by anonymity, but it is not necessary. Common assumptions are that pseudonyms help to deter the "online disinhibition effect" from surfacing and promoting offensive behavior (Suler 2005). Perhaps this "disinhibition effect" played a role in the foundation of 4chan, but the pseudonyms used on r/wtf, r/spacedicks, and r/4chan certainly didn't dissuade it. In fact, it seems that the community itself will determine the level of offensive language that is tolerable and allowed, and for what purposes. This is most evident in r/wtf, which discourages the use of insulting language, but widely promotes the use of those swearword collocates identified in this study as "Disbelief" collocates (holy fuck, what the fuck, etc.). Further examination of the community speech patterns on 4chan would likely show some interesting results, as well as further examination in other anonymous websites, but for now it seems that anonymity doesn't make 4chan offensive or negative, 4chan's use of swearwords and their knowledge of the indexicality of their speech provides a way for anon to recognize another real anon from any number of other site visitors

like redditors, normies, or newfags and form a strong, cohesive group identity. 4chan's reputation emerges not just from their actions online, as these are seemingly connected to inter-group relations. Even their actions on other sites are not unique compared to other shock sites and individual trolls. Instead, the attitudes towards 4chan instead seem more appropriately to be stereotypes generated by a misunderstanding of the goal of the site's group identity. This misunderstanding is even entrenched in 4chan's slang. After all, it's better to be a newfag than a normie it seems.

REFERENCES

Bartlett, J. (2013, October 1). 4chan: The role of anonymity in the meme-generating cesspool of the web. *Wired*. Retrieved from http://www.wired.co.uk/news/archive/2013-10/01/4chan-happy-birthday

Bergs, A. (2006). Analyzing online communication from a social network point of view: Questions, problems, perspectives. *Language@ Internet, 3*. Retrieved from http://www.languageatinternet.org/articles/2006/371

Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., & Vargas, G. G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large Online community. In *ICWSM* (pp. 50-57). Palo Alto, CA: AAAI Press.

Culpeper, J. (2011). *Impoliteness*. Cambridge, England: Cambridge University Press.

Christie, C. (2013). The relevance of taboo language: An analysis of the indexical values of swearwords. *Journal of Pragmatics*, *58*, 152-169.

Dibbell, J. (2010). Radical opacity. *Technology Review*, *113*(5), 82-86.

Eble, C. C. (1996). *Slang & sociability: In-group language among college students*. Chapel Hill: University of North Carolina Press.

Eckert, P. (2006). Communities of practice. *Encyclopedia of language and linguistics*, *2*(2006), 683-685.

Faget, M. (2010). Faget. *Urban Dictionary*. Retrieved from http://www.urbandictionary.com/define.php?term=faget

Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in group decision making communication process and outcome in face-to-face versus computerized conferences. *Human communication research*, *13*(2), 225-252.

Hutto, C. J., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media* (pp. 216-255). Palo Alta, CA: AAAI Press.

Jay, T. (2000). *Why we curse. A Neuro-psycho-social Theory of Speech*. Amsterdam, Netherlands: Benjamins.

Jay, T. (2009). The utility and ubiquity of taboo words. *Perspectives on Psychological Science*, *4*(2), 153-161.

Klima, J. (2015, January 3). Reddit mystery spot (your guide to crazy subreddits): r/wtf. *New Media Rockstars*. Retrieved from http://newmediarockstars.com/2015/01/reddit-mystery-spot-your-guide-to-crazy-subreddits-rwtf/

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., ... Riedl, J. (2012). *Building successful online communities: Evidence-based social design*. Cambridge, MA: Mit Press.

Labov, W. (1968). *The social stratification of English in New York city*. Washington, DC: Center for Applied Linguistics.

McCoy, T. (2014, September 2). 4chan: The 'shock post' site that hosted the private Jennifer Lawrence photos. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/morning-mix/wp/2014/09/02/the-shadowy-world-of-4chan-the-shock-post-site-that-hosted-the-private-jennifer-lawrence-photos/

McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present*. London, England: Routledge.

Mehl, M. R., & Pennebaker, J. W. (2003). The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, *84*(4), 857.

Millen, D. R., & Patterson, J. F. (2003). Identity disclosure and the creation of social capital. In *CHI'03 extended abstracts on Human factors in computing systems* (pp. 720-721). New York, NY: ACM.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of linguistics*, *21*(02), 339-384.

Milroy, L., & Milroy, J. (1992). Social network and social class: Toward an integrated sociolinguistic model. *Language in society*, *21*(01), 1-26.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10* (pp. 79-86). East Stroudsburg, PA: Association for Computational Linguistics.

Pérez-Sabater, C. (2012). A pioneer study on online learning environments following the Common European Framework of Reference for Languages.*Procedia-Social and Behavioral Sciences*, *46*, 1948-1955.

Poole, C. (2010). *The case for anonymity online*. Retrieved from https://www.ted.com/talks/christopher_m00t_poole_the_case_for_anonymity_online?language=en

Rains, S. A. (2007). The impact of anonymity on perceptions of source credibility and influence in computer-mediated group communication: A test of two competing hypotheses. *Communication Research*, *34*(1), 100-125.

Rosenfeld, E. (2012, August 14). Mountain Dew's 'Dub the Dew' online poll goes horribly wrong. *Time*. Retrieved from http://newsfeed.time.com/2012/08/14/mountain-dews-dub-the-dew-online-poll-goes-horribly-wrong/

Santana, A. D. (2014). Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards. *Journalism Practice*, *8*(1), 18-33.

Sharma, V. (2013). *Insult detector model 6.* Retrieved from https://www.kaggle.com/blobs/download/forum-message-attachment-files/320/model6.py

Shuman, P. (2007, July 27). *Fox 11 investigates: 'anonymous'*. Retrieved from http://www.youtube.com/watch?v=DNO6G4ApJQY.

Squires, L. (2010). Enregistering internet language. *Language in Society, 39*(04), 457-492.

Stapleton, K. (2010). Swearing. In M. Locher and S. Graham (eds.), *Interpersonal pragmatics* (Vol. 6, pp. 289-306). Berlin, Germany: Mouton de Gruyter.

Suler, J. (2005). The online disinhibition effect. *International Journal of Applied Psychoanalytic Studies, 2*(2), 184-188.

Tanis, M., & Postmes, T. (2007). Two faces of anonymity: Paradoxical effects of cues to identity in CMC. *Computers in Human Behavior, 23*(2), 955-970.