The Many Shades of Anonymity: Characterizing Anonymous Social Media Content

Denzil Correa[†], **Leandro Araújo Silva**[‡], **Mainack Mondal**[†], **Fabrício Benevenuto**[‡], **Krishna P. Gummadi**[†]

[†] Max Planck Institute for Software Systems (MPI-SWS), Germany

[‡] Federal University of Minas Gerais (UFMG), Brazil

Abstract

Recently, there has been a significant increase in the popularity of anonymous social media sites like Whisper and Secret. Unlike traditional social media sites like Facebook and Twitter, posts on anonymous social media sites are not associated with well-defined user identities or profiles. In this study, our goals are two-fold: (i) to understand the nature (sensitivity, types) of content posted on anonymous social media sites and (ii) to investigate the differences between content posted on anonymous and non-anonymous social media sites like Twitter. To this end, we gather and analyze extensive content traces from Whisper (anonymous) and Twitter (non-anonymous) social media sites. We introduce the notion of *anonymity sensitivity* of a social media post, which captures the extent to which users think the post should be anonymous. We also propose a human annotator based methodology to measure the same for Whisper and Twitter posts. Our analysis reveals that anonymity sensitivity of most whispers (unlike tweets) is not binary. Instead, most whispers exhibit *many shades* or different levels of anonymity. We also find that the linguistic differences between whispers and tweets are so significant that we could train automated classifiers to distinguish between them with reasonable accuracy. Our findings shed light on human behavior in anonymous media systems that lack the notion of an identity and they have important implications for the future designs of such systems.

Introduction

Recently, the Internet has witnessed the emergence of a new type of social media applications and sites called *anonymous social media*. Exemplified by sites like *Whisper* and *Secret*, anonymous social media sites have grown to host millions of users, posting tens of millions of pieces of content that are viewed billions of times every month (Gannes 2013). Compared to traditional social media like Facebook and Twitter, anonymous social media sites make it easier for their users to hide or better protect their (offline or online) identities. For example, Facebook expects every user to post messages from a single online identity (account) that matches their offline identity (e.g., user name). In contrast, Whisper does not associate its users with unique usernames or profile information. This design choice offers Whisper users and their posts a certain degree of anonymity.

While anonymous online forums have been in existence since the early days of the Internet, in the past, such forums were often devoted to certain sensitive topics or issues. In addition, its user population was relatively small and limited to technically sophisticated users with specific concerns or requirements to be anonymous. On the other hand, anonymous social media sites like Whisper[1] and Secret[2] provide a generic and easy-to-use platform for lay users to post their thoughts in relative anonymity. Thus, the advent and rapidly growing adoption of these sites provide us with an opportunity for the first time to investigate how large user populations make use of an anonymous public platform to post content.

In this paper, our goal is to better understand the characteristics of *content* posted on anonymous social media sites. Specifically, we introduce the notion of *anonymity sensitivity* to measure the sensitivity of content posted on such sites. Intuitively, anonymity sensitivity of a message captures the degree to which users prefer to post the message anonymously. We observe that not all users might perceive a given message to be equally sensitive. Some users may prefer to post it anonymously, while others might be comfortable to post the same message non-anonymously, i.e., have their identities associated with the message. Therefore, we propose to measure the anonymity sensitivity of a message by conducting a user study with a large number of participants and computing the fraction of users that would choose to share the content anonymously, if they were required to post the content.

Our analysis in this paper is motivated by the following four high-level research questions about anonymous social media content:

RQ 1 How anonymity sensitive is content posted on anonymous social media sites? How does it compare with content posted on non-anonymous social media sites like Twitter?

RQ 2 What types of content are posted on anonymous social media? Are certain types of content more anonymity sensitive (thus, better suited to be shared over anonymous social media) than others?

[1] `http://whisper.sh`

[2] `https://www.secret.ly`

RQ 3 To what extent do user demographics – such as gender, age, education, and income – affect the perception and measurements of content sensitivity?

RQ 4 Are there significant linguistic differences between content posted on anonymous and non-anonymous social media sites? Can we automatically distinguish between anonymous and non-anonymous media posts by analyzing their linguistic features?

To address the above questions, we gather and analyze extensive traces of content posted on Whisper and Twitter, one of the most popular contemporary anonymous and non-anonymous social media websites, respectively. Specifically, we investigate the short textual content that is contained in both Whisper posts (or whispers) and Twitter posts (or tweets). Our analysis driven by the questions listed above yields many interesting insights into anonymous social media content.

First, we discover many shades of anonymity in whispers. In comparison with tweets, most of which have very low anonymity sensitivity, we find that whispers span the entire range of possible anonymity sensitivity scores (from zero to one). Second, our analysis of the types of content posted in whispers reveals that anonymous social media is used for novel purposes, many of which we did not anticipate. For example, a substantial fraction (≈56.81%) of all whispers relate to public and anonymous *confessions* by users with guilt, shame or embarrassment for things they have done or incidents they participated in the past. Third, our user study reveals statistically significant differences in the measurements of anonymity sensitivity of content taken across different demographic user groups. Specifically, people that are older and have higher education wish to share content more anonymously than others.

Finally, we find clear and significant linguistic differences between whispers and tweets – the former contains more personal, social and informal words than the latter, whispers also express more negative emotions related to sadness and anger, and they communicate more wants, needs, and wishes than tweets. In fact, we show that the linguistic differences between whispers and tweets are so significant that it is possible to build a machine learning system that can distinguish between whispers and tweets by analyzing their textual content.

Our findings help us better understand certain important aspects of human social behavior associated with anonymity. Concretely, we shed light on the type of users that patronize anonymous media sites; the types of content shared by these users and the reasons for their activity. We conclude by discussing the implications of our findings for the design of anonymous social media systems, in which users have weak or no notions of identity.

Related Work

The nature of anonymity in online and offline world and its effect on user behavior has attracted adequate research attention. In this section, we review existing studies on anonymity across three different aspects.

Online user anonymity requirements

Internet users across different age groups and usage interests desire anonymity for various reasons (Kang, Brown, and Kiesler 2013). Users extensively indulge in anonymous content sharing for strategic benefits due to the nature of the content shared and potential social cues (Zhang and Kizilcec 2014). An analysis of anonymous posts on a moms website shows that anonymous content consists of negative emotions – for example, critical content about spouses (Schoenebeck 2013). However, prior studies treat anonymity as binary concept (two-level) viz. anonymous and non-anonymous. In this work, we introduce the concept of *anonymity sensitivity* and transcend the binary to a multi-level concept. To the best of our knowledge, this is the first attempt to quantify *anonymity sensitivity*. We discover that content on anonymous social media has many shades of anonymity or different levels of anonymity.

Anonymity on social media systems

Popular online social media like Twitter and Facebook are non-anonymous where each post is associated with an identity and personally identifying information. The desire for anonymity from users has led to the creation of anonymous content contribution platforms. For example, Slashdot enabled comment anonymity for its users but as of 2008, only 18.6% comments were posted anonymously (Gómez, Kaltenbrunner, and López 2008). Over the years, the scenario has changed with popular anonymous content platforms like *4chan* and */b/* (Bernstein et al. 2011). Recent studies observe a significant growth in anonymity-seeking behavior on online social media (Stutzman, Gross, and Acquisti 2013). Prior work also shows that users create one time use accounts (also called *throwaway*) to post content anonymously (Leavitt 2015). Users who seek the full anonymity experience have found an appropriate medium in anonymous social media sites like Whisper and Secret. These systems allow users to post content without any unique username or associated personally identifiable information. Prior work focus on user engagement and threat attacks on anonymous systems i.e. they try to expose the location of users of such anonymous media (Wang et al. 2014). An analysis of Whisper, a popular anonymous social media, also shows that social links are ephemeral and user engagement primarily is short-lived (Wang et al. 2014). In contrast, our work explores anonymity sensitivity requirements of content across different categories on anonymous social media. We also investigate the effect of personal information viz. demographic variables on anonymity sensitivity.

Anonymity sensitive content characteristics

Social psychology research literature shows that anonymity strongly influences user behavior – online and offline. Humans turn aggressive and violent in situations in an environment that is less constrained by social norms (Zimbardo 1969). Humans also exhibit a disinhibition complex within communications in an anonymous setting (Pinsonneault and Heppel 1997; Suler 2004). In an anonymous environment, people are more likely to shed their hesitation and disclose

more personal information in their communications (Joinson 2001). Similar behavior has also been observed in online anonymous settings. For example, posts on an anonymous mom online forum discuss situations and issues which disregard societal and cultural norms (Schoenebeck 2013). Therefore, we envisage that the content posted on anonymous social media would have unique linguistic signatures. In this study, we discover linguistic characteristics of anonymity media content and contrast it with non-anonymous media content. We leverage these unique lingual signatures to build a machine learning based system to distinguish between anonymous and non-anonymous social media content.

Dataset

In this section, we describe our experimental dataset from a large anonymous social media. First, we will briefly introduce Whisper, a popular anonymous social media sharing site. Next, we outline the details of our experimental dataset

Whisper: anonymous social media sharing site

In this work, our aim is to investigate characteristics of anonymous social media. In order to achieve our objective, we use Whisper, a popular anonymous social media sharing site, as our experimental testbed. Whisper (launched in March 2012) is a mobile application in which users post anonymous messages called "whispers". Whisper is a very popular anonymous social media with more than 2.5B page views, higher than even some popular news websites like CNN (Gannes 2013). Whisper also has 2M+ users and 45% of the users post something every day (Griffith 2013); statistics published by Whisper mention that 70% of their users are women, 4% have age $<$18 years, and most of the Whisper users belong to the age group 17-28.

Whisper users can only post messages via mobile phones, however whisper has a read only web interface. In contrast with traditional social media sites like Twitter, whispers do not contain identifiable information. An initial username is randomly assigned by Whisper, but it is non-persistent i.e., users can change their usernames at any point of time. In addition, multiple users may choose to use the same username. Whisper users also do not have profile pages or information and hence, you can not navigate whispers posted by any particular username. This anonymity property of whispers and the large number of whispers shared per day by the users, make Whisper a very attractive testbed to investigate anonymous social media properties.

Table 1 shows an example of a whisper. Each whisper is overlaid on an image which is randomly chosen or can be provided by the user. A user may also provide location information with whisper at different granularity levels. Each whisper can be favorited (heart) or replied to by another whisper. We see that this particular whisper has 4526 favorites, 390 replies and this whisper originates from Florida, USA.

Whisper experimental dataset

We employ a similar methodology as Wang *et al.* to collect Whisper data (Zhang and Kizilcec 2014). We gather our dataset via the "Latest" section of the Whisper website which contains a stream of publicly posted whispers. Each downloaded whisper contains the text of the whisper, location, timestamp, number of hearts (favorites), number of replies and username. Table 2 shows the overall statistics for our experimental Whisper datasest. Overall, our dataset contains 20.7M whispers with 1.3M usernames and 266,321 locations. We observe that 63.6% of the whispers contain location information. We do not know the total number of Whisper users in our dataset as there is no unique global identifier associated with the user. We recall that Whisper users can change their username at any particular point of time.

Text	My girlfriend lost her eye in an accident and now whenever she texts me and she sends a smiley face she sends ".)" instead of ":)"
Location	Spring Hill, Florida
URL	`https://whisper.sh/w/MjQ4MjE0MDQ1`
Hearts	4526
Replies	390

Table 1: shows an example of a posted whisper.

Time Period	July 1 – November 17 2014
Whispers	20.7M
Usernames	1.3M
Locations	266,321
Whispers with location	63.6%

Table 2: overall statistics for Whisper dataset.

RQ1 : The Many Shades of Anonymity

In this section, we first quantify anonymity sensitivity of content posted on anonymous social media sites and then compare it to that of non-anonymous social media sites.

Measuring anonymity sensitivity of content

We observe that not all users might perceive a piece of content to be equally anonymity sensitive. In order to quantify this variation of anonymity sensitivity of a content, we setup the following Amazon Mechanical Turk (AMT) experiment: we pick 500 random whispers from our crawled Whisper dataset and 500 random publicly available tweets from Twitter's Streaming API sample (Twitter Streaming API 2015). We ask 100 AMT workers to annotate each of these 500 tweets and 500 whispers as anonymous or non-anonymous. We do not reveal the origin of the tweets or whispers to AMT workers. In order to ensure annotation quality, we choose only master AMT workers and compensate approximately 10 USD/hour per annotator. In order to take further precautions against noisy annotations, we consider a part of the data as gold standard. We observe that 14 out of the 100 AMT workers fail our gold standard test. We discard these 14 workers and consider annotations only from the remaining 86 AMT workers. These data hygiene checks help in data quality assurance and noise elimination.

In our experiment, AMT worker representation is one potential concern which may lead to a sampling bias. In other words, worker annotations might not reflect the Internet population. In order to address this concern, we collect demographic information from AMT workers. The presence of a demographic information page is made explicit in our survey description and all demographic fields are optional. However, the demographic information is requested at the end of the survey and is not displayed to the user until the annotations are complete. We also state that the demographic data will be only used for research purposes and we would not disclose personally identifiable information. Since our demographic information setup is transparent, clear and optional – the demographic variables collected are unlikely to have noise. We find that 88% of our annotators are from US with 51.2% females. The male-to-female ratio is 0.95, approximately equivalent to the US population (Howden and Meyer 2010). The age demographic of our AMT workers vary between 18 to 74 years. The median age is 32 years and is slightly lower than that the median age of the US population (Howden and Meyer 2010). The AMT annotators have a wide income range from under 10,000 USD to more than 100,000 USD per annum with a median annual income of 27,000 USD. The information demonstrates that our annotators have varied and are consistent with prior studies of AMT workers (Ross et al. 2009).

In our annotated data, each text (tweet or whisper) is marked by 86 AMT workers as anonymous or non-anonymous. The fraction of AMT workers who annotate each text as anonymous is a probabilistic estimate of the fraction of users that would consider the text as anonymous. We call this probabilistic estimate the Anonymity Sensitivity Score (or *AS Score*) for that particular text. Formally, the *AS Score* for a given piece of text is the probability that users would consider this text as anonymous. Formally for an AMT worker AMT_j annotating a text t_i,

$$Score_{t_i}^{AMT_j} = \begin{cases} 0 & \text{when } AMT_j \text{ marks } t_i \text{ non-anonymous} \\ 1 & \text{when } AMT_j \text{ marks } t_i \text{ anonymous} \end{cases}$$

$$\text{AS Score}_{t_i} = \frac{\sum_{j=1}^{86} Score_{t_i}^{AMT_j}}{86}$$

Table 3 shows examples of messages from our AMT experiment for different *AS Scores*. We can see that the message *"Is it to late to join a beer softball league?"* was labeled as anonymous by 43 AMT workers and as non-anonymous by the remaining 43 workers, giving an *AS Score* of 0.5 to this text. We note that pieces of content with higher *AS Scores*, i.e. which higher number of AMT workers annotate as anonymity sensitive, tend to be more controversial and intuitively require more anonymity. Behavioral studies in psychology have also shown that anonymity leads people to reveal sensitive content (Suler 2004).

Anonymous versus non-anonymous media

In order to compare anonymity sensitivity of content on anonymous and non-anonymous media, we inspect *AS Score* probability distributions of whispers and tweets. Figure 1 shows the cumulative and probability distributions of *AS Score* for anonymous and non-anonymous media.

AS Score	Example
0.0	absolutely roastin today!!!! im dying lol x
0.1	people really scare me sometimes...
0.2	Benefits of long-distance relationships?
0.3	People fear me cause i am a bad influence.
0.4	Looking for a cowboy! <3
0.5	Is it to late to join a beer softball league?
0.6	I love it when a female bits my neak
0.7	hey ebony girls i am waiting
0.8	gay guy wanting some guy for nsa hook up
0.9	Free the nipples
1.0	I suck at sexting! 19y female.

Table 3: shows example messages from our AMT experiment with different AS Scores.

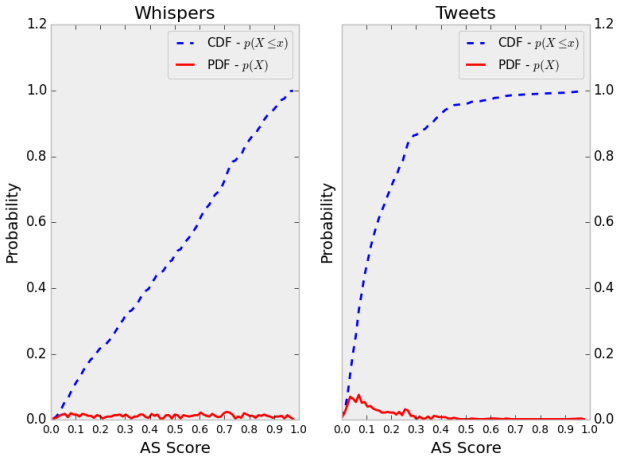

Figure 1: shows the cumulative distribution function and probability distribution function of *AS-Score* for whispers and tweets.

We notice that ≈16% of whispers have an *AS Score* ≥0.8 while only 1% of tweets exceed the same level. Hence we conclude, anonymous social media contains more anonymous sensitive content than non-anonymous social media. Moreover, the *AS Score* distribution on Twitter tends to be concentrated at ≤0.3 levels (86.4% of tweets have a score of ≤0.3), which signifies that a vast majority of Twitter messages are non-anonymous or were considered as anonymous only by a small fraction of the AMT workers. On the other hand, we see that the probability distribution of *AS Scores* on Whisper are uniform. Therefore, content posted on Whisper has many levels (or shades) of anonymity. In addition, we also calculated the inter-annotator agreement between the 86 AMT annotators using Fleiss Kappa. We find that agreement is -0.015 which indicates a poor rating and therefore, shows that anonymity sensitivity varies between individuals. Contemporary research treats anonymity as a binary concept with systems designed to cater to either anony-

mous or non-anonymous content. However, research in behavioral psychology suggests that anonymity is *subjective* in nature (Pinsonneault and Heppel 1997). Our experiment also shows that anonymous sensitivity transcends binary notions and has different levels or many shades.

Summary

In this section, we first propose a methodology to measure anonymity sensitivity of content and then apply it to analyze the anonymity sensitivity of whispers and tweets. We discover that while most tweets are not anonymity sensitive, whispers have many shades or levels of anonymity, i.e., their anonymity sensitivity scores span the entire range of possible values from zero to one. Thus, while anonymous media content is significantly more sensitive (overall) than non-anonymous media content, not all anonymous media content is equally sensitive.

RQ2 : Content Categories and Anonymity Sensitivity

In this section, we categorize anonymous social media content and investigate anonymity sensitivity of content based on its category. We discover anonymity sensitivity scores vary significantly across the different types (categories) of content.

Categories of anonymous media content

We observe that Whisper automatically classifies each posted message into one or more content category. 15.78% of whispers are not classified in any category. Overall, Whisper has 33 existing categories (Figure 2) and we manually inspected 100 messages for each of the 33 categories. Our manual analysis revealed some severe limitations of existing Whisper categories and inspired us to propose a new set of categories. We validate the utility of our new set of categories via an Amazon Mechanical Turk(AMT) experiment.

In existing Whisper categories, we discover that most messages belonging to *Politics*, *DIY and Home*, *Work* and *Sports* are not actually related to their mapped categories. Therefore, we discard these categories from our consideration. We also find that whispers posted in certain categories are very similar to each other and hence, we merged these categories. For example, *Family* and *Relationships* could be merged as *Relationships*. Hence, we propose 9 categories that adequately represent anonymous content. Table 4 shows the proposed content categories, their corresponding Whisper categories and examples of whispers in each category.

We validate that our proposed 9 content categories are sufficient to cover a vast majority of content posted on Whisper via an AMT experiment. We randomly select the same 500 whispers we used in the AMT experiment from the previous section and request feedback from 5 AMT annotators. We choose master AMT workers from US who are expert annotators to assure annotation quality. Each annotator is paid approximately 10 USD/hour. Each AMT annotator is asked to categorize all the 500 whispers into one or more of the 9 categories in Table 4. We also provide an *Other* label with a text

Confessions	LGBTQ	DIY & Home
Love & Romance	Meet Up	Fashion
LOL	Health & Wellness	Travel
OMG	Drugs & Alcohol	Events & Places
WTF	Money	Science & Tech
Advice	School	Food
Q&A	Work	Politics
NSFW	Military	Sports
Family	Faith	Animals & Pets
Parenting	Celebrities & Culture	News
Relationships	Entertainment & Arts	Tattoos & Piercings

Figure 2: shows the 33 high level content categories designed by Whisper

field in case the AMT annotator feels that none of the 9 categories represent the whisper. Figure 3 shows the cumulative distribution function of AMT annotator agreements on category labels. We see that 3 AMT annotators (majority vote) agree that 93.8% Whispers are captured by the 9 provided categories in the survey. Hence, we conclude that the proposed 9 categories are sufficient to cover the content posted on Whisper.

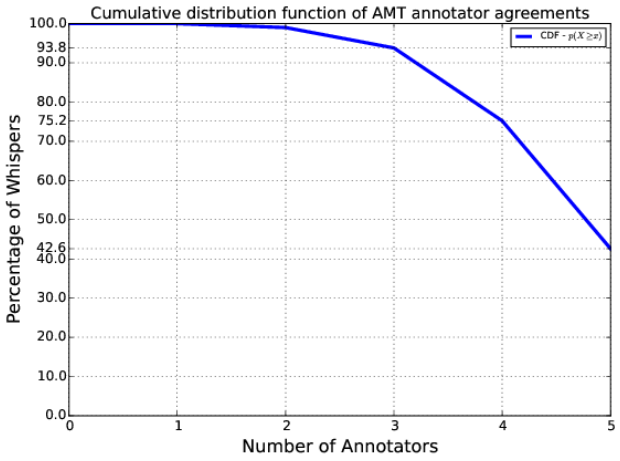

Figure 3: shows the cumulative distribution function of AMT annotator agreements on category labels. 3 AMT annotators (majority vote) agree that 93.8% Whispers are captured by the 9 provided categories.

Table 5 shows the distribution of whispers in the proposed 9 content categories as agreed upon by 3 AMT annotators (477 whispers). We see that *Confessions*, *Relationships*, *Meetup* and *QnA/Advice* are the top categories, accounting together for 90.97% of all the texts. 1.68% of whispers were categorized as *Other*. We analyze the text data provided by AMT annotators via the *Other* label and do not find any significant categories. Therefore, we conclude that people use anonymous social media to largely post content about *Confessions*, *Relationships*, *Meetup* and *QnA/Advice*.

Selected Category	Corresponding Whisper Category	Examples
Meetup	Meetup	any ladies 18+ want a new snap friend? 30/m here...
Confessions	Confessions	For all of my college life, the only makeup I ever had was the makeup I stole from house parties.
Relationships	Relationships, Family, Parenting	My mother wants me converse more with her. Yet when I try, the discussion turns to me being fat, useless and pathetic. She wonders why I don't speak to her.
NSFW	NSFW	boobs <3
LOL	LOL	Each time I see a whisper on the popular page I think damn why didn't I think of that
QnA/Advice	QnA/Advice	i had unprotected sex while i was on my period, and i forgot to take my pills ... should i be concerned? i still amaze myself with all the wrong decisions i make
Health and Wellness	Health and Wellness	I've been struggling with eating and disorders for 10 months now and no one knows
LGBTQ	LGBTQ	I think I'm bi. I like men. But I also have interest in grlz. I check grlz out & have a desire 2 be wit them sexually. Yes I kissed a girl & I liked it. I did not say that cuz of the song, btw.
Drugs and Alcohol, Black Markets	Drugs and Alcohol, Miscellaneous	Male escort in London requires female assistant for a few jobs in the next 2 weeks.Cash in hand

Table 4: shows the proposed content categories, corresponding Whisper category and examples of whispers in each category.

Category	Percentage of Whispers
Confessions	56.81
Relationships	24.74
Meetup	17.82
QnA/Advice	14.47
NSFW	7.13
LGBTQ	5.03
Health and Wellness	5.03
LOL	4.82
Drugs and Alcohol, Black Markets	2.52
Others	1.68

Table 5: shows the distribution of whispers agreed upon by 3 AMT annotators (477 whispers) in the proposed 9 content categories.

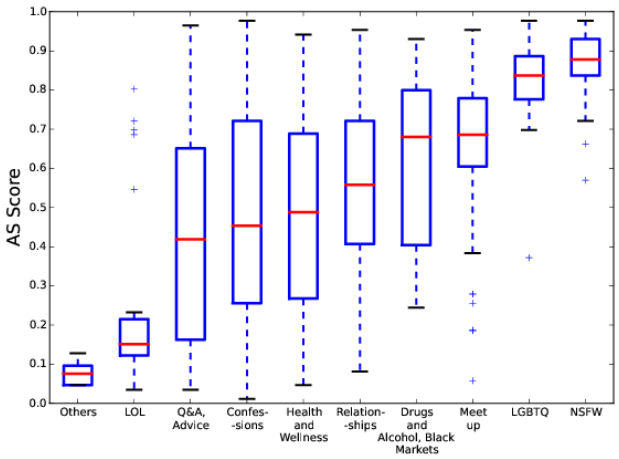

Figure 4: shows the distribution of *Anonymity Sensitivity Score* for all categories via box-and-whisker plot. *NSFW*, *Relationships* and *LGBTQ* categories contains highly sensitive content while *LOL* has low content sensitivity.

Does anonymity sensitivity vary across categories?

In this section, we examine the differences in anonymity sensitivity of content across categories. Previously, we setup two different AMT experiments using the same set of 500 random whispers. The first experiment helped us quantify the *anonymity sensitivity* of each whisper while the second experiment classified each whisper into a content category. Now, we combine the results from both the aforementioned AMT experiments and compute the *Anonymity Sensitivity Score* for each category. Figure 4 shows the distribution of this score for all categories via a box-and-whisker plot.

We observe that categories like *NSFW*, *LGBTQ*, *Meet Up* and *Drugs/Alcohol/Black Markets* have a very high *AS Score* (median $\geq\approx 0.7$) while LOL has a low *AS Score* (median $\leq$0.2). We also see categories like *Health and Wellness*, *QnA/Advice*, *Relationships* and *Confessions* have moderate *AS Score* (median $\geq$0.4 and $\leq$0.55). Therefore, *NSFW*, *Meet Up*, *Drugs/Alcohol/Black Markets* and *LGBTQ* categories contains highly sensitive content while *LOL* has low content sensitivity. Overall, we notice that content sensitivity varies significantly across categories and different categories contain content with different levels of anonymity sensitivity.

Summary

We analyze different types of content posted on anonymous social media sites and find that the a large fraction of content posts on the correspond to *Confessions*, *Relationships*, *Meetup* and *QnA/Advice*. We examine the anonymous sensitivity of content in each category and discover that content in certain categories (*NSFW* and *LGBTQ*) is significantly more anonymity sensitive than others. Given the considerably different levels of anonymity desired for content in different categories, in the future, it might be worth investigating system designs that offer different levels of anonymity protections and guarantees.

RQ3 : User Demographics and Anonymity Sensitivity

Previously, we observed variance of anonymity sensitivity by content type. In this section, we investigate how the perception and measurement of anonymity sensitivity of a given content varies across different user demographics.

Demographic data collection

We recall that we collect demographic information from AMT workers. In addition, our demographic collection process follows best practices including transparency and therefore, eliminates significant potential bias. We use the same set of 477 random sampled whispers with 3 AMT annotator agreements from the previous section as our experimental dataset. For each AMT worker, we gathered the following demographic information: *Nationality, Country of Residence, Gender, Age, Education, Employment Status, Income, Political View, Race, Marital Status*. We note that each demographic variable was asked as a well phrased question using best survey practices. The options for each variable also consists of standard survey options for the respective questions. We are unable to report the entire survey due to space constraints.

We inspect each demographic variable of users across different content categories and *AS Score* values. For every demographic variable, we divide AMT workers into multiple categories based on their responses. We then analyze the *AS Score* distributions between these demographic variable responses. We use Mann—Whitney U test to determine statistical significance ($p < 0.05$) in *AS Score* distributions across each response categories (Mann and Whitney 1947). The distributions that present statistically significant differences are inferred as demographic effects on anonymity sensitivity. Prior work has found that demographic information like age, gender and education have an effect on an individual's privacy concerns (Associates and Westin 1993). In the same line, we examine the relationship of age, gender and education on anonymity sensitivity.

Effect of gender: We find no statistically significant difference between AS scores assigned by these two groups. In contrast, previous studies show that females are more privacy sensitive than males (Miyazaki and Fernandez 2001). We further investigate the *AS score* distributions for gender across 9 anonymous content categories we defined earlier. Our analysis shows that for the *NSFW* category, females are significantly more anonymous sensitive than males. We conclude that gender affects anonymity sensitivity only if the content is highly *NSFW*.

Effect of age: We divide our AMT workers into two age categories : younger users with age between 18 to 34 years and older users with age between 35 to 74 years. We observe a statistically significant difference in *AS scores*, where older workers are more anonymity sensitive than younger workers. We further check the AS score distributions across content categories. We find that older workers are significantly more anonymity sensitive to messages that belong to *Relationships*, *Confessions* and *NSFW* content categories. Therefore, we conclude that age increases anonymity sensitivity if it belongs to the aforementioned three categories.

Effect of education: Approximately, 15.1% of our workers did not have a college education. Therefore, we divide workers into two groups: College educated users and non-college educated users. We find that college educated users are more (statistically significant) anonymity sensitive than non-college educated users. We further analyze AS score distributions for both groups across content categories. We observe that college educated users are significantly more anonymity sensitive in almost all the categories apart from *LOL* and *Drugs and Alcohol*. Therefore, we conclude that college education increases anonymity sensitivity regardless of content.

Effect of income: We categorize our workers according to their annual income in two groups: workers whose annual income is less than the median income of the US population (30,000 USD) and those above it (Howden and Meyer 2010). The AS score distribution of lower income groups are higher (statistically significant) than that of higher income group. However, we find that 23.5% of workers from the higher income category are non-college educated while that of lower income group is only 10.2%. In addition, 42.9% of workers from lower income group are currently enrolled in a college. We notice that despite college education the workers do not earn enough to be categorized into the high income group. Therefore, we conclude college education makes workers more anonymity sensitive despite their income category.

Summary

We observe significant differences in anonymity sensitivity scores of a content measured across different user demographics. We find that college education has a significant (increasing) effect on anonymity sensitivity despite income variations, while age and gender predominantly affect certain anonymous content categories. Our findings hint at the target demographics (potential users) for anonymous social media systems.

RQ4 : Linguistic Analysis of Anonymous Media Content

In the previous sections, we observed that anonymity sensitive content varies according to content type and user demographics. In this section, we investigate the linguistic characteristics of content posted on anonymous social media. First, we mine textual data to find out unique linguistic signatures of anonymous media content. Second, we leverage the linguistic characteristics to build a system to differentiate between anonymous and non-anonymous media content.

Linguistic characteristics of anonymous content

Our goal is to understand the distinguishing linguistic characteristics of content posted on anonymous social media. Therefore, we contrast and compare the textual content on anonymous social media to that of non-anonymous social media. We use Twitter as our non-anonymous media. We consider a random sample of 100,000 tweets and 100,000 whispers for our experiment. The whispers were drawn from

the period between July 1, 2014 and November 17, 2014. The tweets were drawn from Twitter's Streaming API sample for the same duration (Twitter Streaming API 2015). We extract stemmed (Porter Stemmer) unigrams from both whispers and tweets including of the stopwords. We then categorize these unigrams into different dictionaries provided by LIWC (Tausczik and Pennebaker 2010). LIWC is a hierarchical linguistic lexicon that classifies words into 70 meaningful psychological categories. We calculate the percentage of tweets and whispers which belong to each LIWC dictionary. We also calculate the whispers to tweets ratio for each dictionary. Since our aim is to investigate linguistic characteristics of anonymous media content, we focus on the dictionaries which have both significant percentage and a high Whisper to Twitter ratio. Table 6 shows the linguistically significant LIWC dictionaries. [3]

Whispers are more personal, social and informal than tweets – Whispers have a significantly high presence of *1st Person Singular Pronouns* than tweets. Prior studies indicate this is a psycholinguistic characteristic of personal and informal content (Tausczik and Pennebaker 2010). We also observe a significant presence of *Humans* dictionary which indicates that users talk about their social life interactions. Whispers also contain a significant percentage of words which belong to the *sexual* dictionary, being 2.2x times more than tweets. On the other hand, Twitter has a significant presence of *Work* and *Money* dictionaries which indicates a more formal environment. Therefore, anonymity on social media exhibits the *disinhibition* effect where users can shed their inhibitions and post personal sensitive content about their daily lives in informal language (Joinson 2001; Suler 2004).

Whispers communicate more negative emotions due to sadness and anger than tweets – Whispers contain a high percentage of *Positive Emotion* and *Negative Emotion* dictionaries. However, in comparison to Twitter *Negative Emotion* words appear ≈1.8x times on Whisper. Therefore, anonymous social media is an emotionally charged environment and most emotions are negative. Whispers also have significant high ratios of *Sadness* and *Anger* LIWC dictionaries, which suggests that an anonymous environment provides an outlet for users to vent out frustration, display states of anger and sadness.

Whispers express more wants, needs and wishes – Whispers also have a high percentage and ratio of words that belong to the *Discrepancy* dictionary. This dictionary consists of words like want, need and wish. We look into the structure of these words in whispers using the Word Tree visualization (Wattenberg and Viégas 2008). Figure 5 shows the Word Tree visualization for word "want" in the *Discrepancy* dictionary.[4] We notice that users express a lot of desires and life needs like someone to talk and emotional distress calls. We observe similar pattern for other words from the *Discrepancy* dictionary like "need" and "wish". Based on our findings, we posit that anonymity provides a congenial platform for users to express their *wants*, *needs* and *wishes*.

[3] We experimented with multiple random samples and observe similar results.

[4] http://www.jasondavies.com/wordtree/

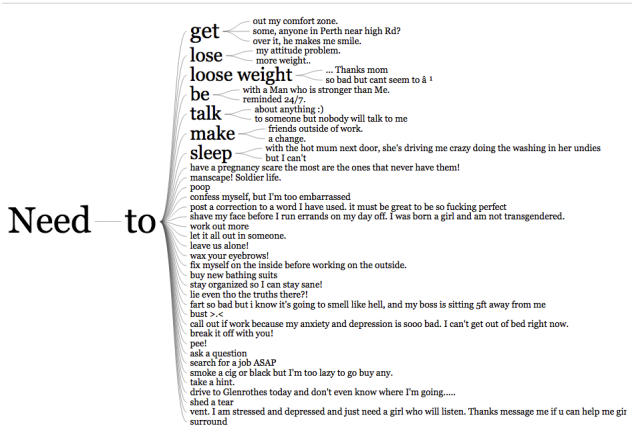

Figure 5: Word Tree visualizations for searches of the word "need to" in the *Discrepancy* LIWC dictionary for all whispers. The visualization shows phrases that branch off from the selected words across all the text in our dataset. A larger font-size means that the word occurs more often.

Automated detection of anonymous media content

We discover that whispers (anonymous media content) has certain unique linguistic characteristics. We leverage these linguistic characteristics to develop a predictive model to distinguish between whispers and tweets (non-anonymous media content). An automated system to detect whispers is a challenging task due to subjective nature of anonymity. We frame our problem as a binary classification problem - anonymous versus non-anonymous media content. Table 7 summarizes the experimental setup for our machine learning framework.

Dataset	200,000 texts (random)
Anonymous	100,000 Whispers
Non-Anonymous	100,000 Tweets
Classifier	Logistic Regression (l2-regularization)
Cross Validation	10-fold
Learning Rate(C)	1.0

Table 7: experimental setup for our prediction task.

We use 100,000 whispers and 100,000 tweets sample from the previous section as our experimental dataset. Similar to linguistic analysis, we consider whispers as anonymous content while tweets as non-anonymous content. We formulate a supervised machine learning framework with the LIWC dictionaries as our feature set. We use Logistic Regression with a $L2$-norm for regularization to prevent our classifier from bias or variance. We perform 10-fold cross validation to avoid data overfit and hence, ensure generalization abilities of the classifier. We evaluate the performance of our classifier on standard evaluation metrics *Precision*, *Recall*, *F1* and *Accuracy*. Table 8 shows the evaluation of our classifier.

Our predictive model obtains an accuracy of 73.12% and an F1 score of 72.88%. We see that our classifier is able to detect whispers from tweets using psycholinguistic features

LIWC Dictionary	Whisper to Twitter Ratio	Whispers%	Tweets%	Example Words	LIWC Category
1st Person singular Pronoun	1.85	72.56	39.2	I, me, mine	Linguistic
Humans	3.71	26.09	7.03	Adult, baby, boy	Psychological – Social
Money	0.56	4.07	7.32	Audit, cash, owe	Personal Concerns
Work	0.75	8.33	11.15	Job, majors, xerox	Personal Concerns
Positive emotion	1.24	42.16	33.93	Love, nice, sweet	Psychological – Affective
Negative Emotion	1.79	28.79	16.07	Hurt, ugly, nasty	Psychological – Affective
Sadness	1.98	6.47	3.26	rying, grief, sad	Psychological – Affective
Anger	1.62	14.12	8.71	Hate, Kill, annoyed	Psychological – Affective
Discrepancy	2.29	35.86	15.66	wish, want, need	Psychological – Cognitive
Sexual	2.18	19.75	9.07	Horny, love, incest	Psychological – Biological

Table 6: linguistically significant LIWC dictionaries.

Precision	73.53% +/- 0.006
Recall	72.26% +/- 0.008
F1	72.88% +/- 0.0009
Accuracy	73.12% +/- 0.001

Table 8: evaluation of our classification task.

(LIWC) with reasonable accuracy. We further analyze the feature set to understand the discriminatory powers to predict whispers. Logistic Regression provides coefficients for each feature which indicates its discriminatory power. Figure 6 shows the top significant features for our prediction task.

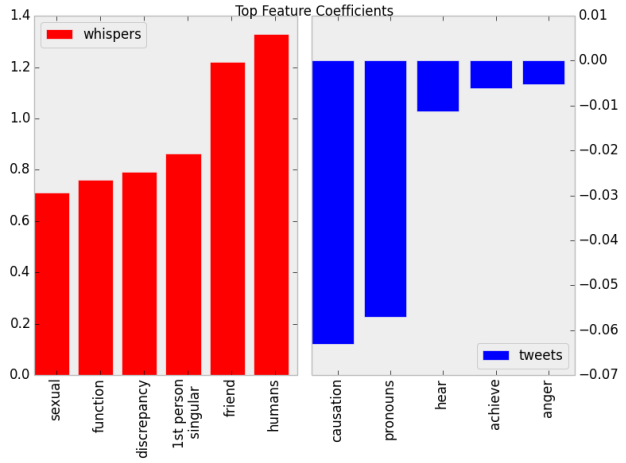

Figure 6: top significant features for our prediction task.

We observe that LIWC categories like *humans*, *friends*, *1st person singular pronoun*, *discrepancy* and *sexual* are most important features to predict whispers. Therefore, we conclude that whispers (anonymous media content) have different psycholinguistic textual properties than tweets (non-anonymous media content).

Summary

We find that whispers have a distinctive linguistic signature compared to tweets. Whispers' content is personal in nature, filled with negative emotions and expresses *wants*, *needs* and *wishes*. We leverage the predictive power of these psycholinguistic categories to build an automated system to differentiate whispers and tweets. Our findings suggest that it might be possible to build classifiers that can automatically distinguish between broad classes of anonymity sensitive and anonymity non-sensitive content. Such classifiers can have tremendous applications in protecting privacy of users in the future.

Concluding Discussion

In this paper, we take the first step towards understanding the nature (sensitivity, types) of content posted on anonymous social media sites. We present a detailed characterization of content posted on Whisper and contrast it with the content posted on Twitter (a non-anonymous social media site). Our analysis revolves around anonymity sensitivity measurement of whispers and yields several interesting insights. We find that whispers span the entire range of possible anonymity sensitivity scores from zero to one and that they are used for various purposes, including public anonymous confessions of guilt, shame or embarrassment over past actions. We also that discover that anonymity sensitivity of content varies across user demographics. In addition, we demonstrate that it is possible to leverage the linguistic differences between whispers and tweets to build an automated system that can distinguish anonymous media content from non-anonymous media content with reasonable accuracy.

Our findings have important implications for security researchers and systems designers analyzing and building anonymous social media systems. Security researchers reason about the anonymity offered by a system in terms of guarantees a system offers. That is, a system either offers strong guarantees to protect user anonymity or none at all. In terms of such an evaluation, most anonymous social media systems (including Whisper) fail to offer strong guarantees for user anonymity protection. Our study shows that in practice, the different content instances posted on these sites have different levels of anonymity sensitivity. Our observations suggest that not all content posted on Whisper needs similar levels of anonymity protection and guarantees. We call for an investigation of anonymity preserving system designs in the future that are specifically optimized for content with a certain level of anonymity sensitivity.

Our results also shed light on certain aspects of human behavior associated with anonymity. For example, although social psychology research literature shows that anonymity strongly influences user behavior (Zimbardo 1969; Pinsonneault and Heppel 1997; Suler 2004), these efforts suggest that individuals turn more aggressive and exhibit disinhibition in an anonymous environment. The characterization of content of an online widely used anonymous system shows that users also use anonymity to express their *wants*, *needs*, and *wishes*.

Finally, our study shows the feasibility to differentiate whispers from tweets by linguistic analysis of the content. We also build an automated system using a machine learning framework to reasonably differentiate whispers and tweets. In the future, we plan to investigate the potential for leveraging the ability to distinguish anonymous media content from non-anonymous media content to design automated anonymity advisors that alert users before they post highly anonymity sensitive messages.

Acknowledgments

This research was supported in part by a grant from the Indo-German Max Planck Centre for Computer Science (IMPECS). Fabrício Benevenuto is supported by grants from CNPq, CAPES, and Fapemig.

References

Associates, L. H. ., and Westin, A. F. 1993. *Health Information Privacy Survey*. Equifax Inc.

Bernstein, M. S.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. G. 2011. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*.

Gannes, L. 2013. On making our digital lives more real. http://allthingsd.com/20130802/im-so-over-oversharing-on-making-our-digital-lives-more-real/.

Gómez, V.; Kaltenbrunner, A.; and López, V. 2008. Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the 17th International Conference on World Wide Web*, 645–654. ACM.

Griffith, E. 2013. With 2 million users, "secrets app" whisper launches on android. http://pando.com/2013/05/16/with-2-million-users-secrets-app-whisper-launches-on-android/.

Howden, L. M., and Meyer, J. A. 2010. Age and sex composition: 2010. *2010 Census Briefs, US Department of Commerce, Economics and Statistics Administration. US CENSUS BUREAU*.

Joinson, A. N. 2001. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology* 31(2):177–192.

Kang, R.; Brown, S.; and Kiesler, S. 2013. Why do people seek anonymity on the internet?: informing policy and design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2657–2666. ACM.

Leavitt, A. 2015. "this is a throwaway account": Temporary technical identities and perceptions of anonymity in a massive online community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, 317–327. New York, NY, USA: ACM.

Mann, H., and Whitney, D. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 50–60.

Miyazaki, A. D., and Fernandez, A. 2001. Consumer perceptions of privacy and security risks for online shopping. *Journal of Consumer Affairs* 35(1):27–44.

Pinsonneault, A., and Heppel, N. 1997. Anonymity in group support systems research: A new conceptualization, measure, and contingency framework. *Journal of Management Information Systems* 89–108.

Ross, J.; Zaldivar, A.; Irani, L.; and Tomlinson, B. 2009. Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep*.

Schoenebeck, S. Y. 2013. The secret life of online moms: Anonymity and disinhibition on youbemom. com. In *ICWSM*.

Stutzman, F.; Gross, R.; and Acquisti, A. 2013. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality* 4(2):2.

Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7(3):321–326.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

2015. Twitter Streaming API. `https://dev.twitter.com/streaming/reference/get/statuses/sample`.

Wang, G.; Wang, B.; Wang, T.; Nika, A.; Zheng, H.; and Zhao, B. Y. 2014. Whispers in the dark: Analyzing an anonymous social network. In *Proceedings of the 2014 conference on Internet measurement conference*. ACM.

Wattenberg, M., and Viégas, F. B. 2008. The word tree, an interactive visual concordance. *Visualization and Computer Graphics, IEEE Transactions on* 14(6):1221–1228.

Zhang, K., and Kizilcec, R. F. 2014. Anonymity in social media: Effects of content controversiality and social endorsement on sharing behavior. In *ICWSM*.

Zimbardo, P. G. 1969. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation*.