Whispers in the Dark: Analysis of an Anonymous Social Network

Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng, Ben Y. Zhao
Department of Computer Science, UC Santa Barbara
{gangw, bolunwang, tianyi, anika, htzheng, ravenben}@cs.ucsb.edu

ABSTRACT

Social interactions and interpersonal communication has undergone significant changes in recent years. Increasing awareness of privacy issues and events such as the Snowden disclosures have led to the rapid growth of a new generation of anonymous social networks and messaging applications. By removing traditional concepts of strong identities and social links, these services encourage communication between strangers, and allow users to express themselves without fear of bullying or retaliation.

Despite millions of users and billions of monthly page views, there is little empirical analysis of how services like *Whisper* have changed the shape and content of social interactions. In this paper, we present results of the first large-scale empirical study of an anonymous social network, using a complete 3-month trace of the Whisper network covering 24 million whispers written by more than 1 million unique users. We seek to understand how anonymity and the lack of social links affect user behavior. We analyze Whisper from a number of perspectives, including the structure of user interactions in the absence of persistent social links, user engagement and network stickiness over time, and content moderation in a network with minimal user accountability. Finally, we identify and test an attack that exposes Whisper users to detailed location tracking. We have notified Whisper and they have taken steps to address the problem.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; K.6 [**Management of Computing and Information Systems**]: Security and Protection

General Terms

Measurement; Design; Security

Keywords

Anonymous Social Networks; Graphs; User Engagement; Privacy

IMC'14, November 5–7, 2014, Vancouver, BC, Canada.

http://dx.doi.org/10.1145/2663716.2663728 .

1. INTRODUCTION

Over the last decade, online social networks (OSNs) such as Facebook, LinkedIn, and Twitter have revolutionized the way we communicate. By formalizing our offline social relationships into digital form, these networks have greatly expanded our capacity for social interactions, both in volume and frequency.

Yet the industry landscape is changing. Content posted on Facebook is now commonly used to vet job candidates, support divorce litigation, and terminate employees. In addition, studies have observed a significant growth in privacy-seeking behavior, even despite changes in social networks to encourage broader information sharing [34]. Finally, these trends have only been accelerated by recent revelations following the Snowden disclosures, with numerous headlines reminding Internet users that their online behavior is under constant scrutiny by NSA and other entities.

All these have contributed to the rapid rise of a new wave of privacy-preserving communication and social networking tools. These fast-growing services are pseudo-anonymous messaging mobile applications: *SnapChat* made headlines for ensuring that photos self-destruct in a few seconds; *Whisper* allows users to anonymously post their thoughts to a public audience; and *Secret* allows users to share content with friends without revealing their own identity. This is just the tip of the iceberg, as many similar services are arriving with increasing frequency, *e.g.*, Tinder, Yik-yak, Wickr.

The anonymous nature of these communication tools has drawn both strong supporters as well as vocal critics. Supporters believe that they provide valuable outlets for whistleblowers avoiding prosecution, and allow users to express themselves without fear of bullying or abuse [40, 41]. Critics argue that the lack of accountability in these networks enables and encourages negative discourse, including personal attacks, threats, and rumor spreading [2, 4]. Yet all parties agree that these tools have had a dramatic impact on how users interact and communicate.

In this paper, we describe our experience and findings in our effort to study pseudo-anonymous social networks, through a detailed measurement and analysis of *Whisper*[1]. Whisper is a mobile app that allows users to post and reply to public messages on top of an image (*e.g.* Internet memes), all using anonymous user identifiers. Whisper does not associate any personal identifiable information with user IDs, does not archive any user history, and does not support persistent social links between users. These design choices are the polar opposite of those in networks such as Facebook. Yet they have made Whisper one of the most popular new social networks, with more than 3 billion page views per month[2]. As our

[1]Our study was reviewed and approved by UCSB Office of Research IRB under protocol `#COMS-ZH-YA-010-6N`.

[2]To the best of our knowledge, there is no publicly available data on user counts in Whisper.

working dataset, we captured 100% of the Whisper data stream for a 3-month period starting in February 2014, including more than 24 million whispers and replies written by more than 1 million unique users.

We focus our study on the net impact of anonymity in Whisper, compared to traditional social media with verified identities and social links. Given the large differences between Whisper and current leaders such as Facebook and LinkedIn, our analysis can have significant implications on future infrastructures for social networks, issues of user privacy in messaging networks, and our understanding of social behavior. More concretely, our study also sheds light on the long-term sustainability of anonymous communication networks, given the removal of persistent social links, often considered key to the "stickiness" of today's networks.

Our analysis provides several key findings.

- First, we seek to understand user interactions in the absence of social links. We build interaction graphs and compare them with those of traditional social networks like Twitter and Facebook. Not surprisingly, we find that user communication patterns show high dispersion, low clustering, significantly different from prior systems. Per user, we observe that "friends" are highly ephemeral, and strong, long-term friendships are rare.
- Second, our study of user activity over time shows that a constant stream of new users contribute significantly to content generation, and users bifurcate clearly into short-lived (1-2 days) and long-term users. We demonstrate that users can be accurately classified into either group by applying ML techniques to only 1 week's worth of activity history.
- Third, we study the question of abusive content through analysis of "deleted whispers." We show that most deleted whispers focus on adult content, and Whisper's moderation team usually deletes offensive whispers within a short time after initial posting.
- Finally, we identified a significant attack that exposes current Whisper users to detailed location tracking. We describe the attack in detail and our experiments. Note that we have already notified Whisper of this vulnerability, and they are taking active steps to mitigate the problem.

To the best of our knowledge, our work is the first detailed study on Whisper and pseudo-anonymous messaging systems as a group. Their rapid user growth on mobile platforms suggests they may offer a real challenge to today's established OSNs. We believe our initial work sheds light on these systems as new platforms for interpersonal communication, and provides insight into designs for network infrastructures to support Whisper and similar services.

2. BACKGROUND AND GOALS

In this section, we briefly describe background information about the Whisper network, followed by a high level summary of the goals of our study.

2.1 Background: the Whisper Network

Whisper.sh is a two-year old smartphone app that has become a leader in a new wave of pseudo-anonymous messaging and social communication services, including *Snapchat*, *Secret*, *Tinder*, *Yikyak*, *Ether* and *Wickr*. While detailed functionality may vary, these services generally provide ways for users to make statements, share secrets or gossip, all while remaining anonymous and untrackable.

As a mobile-only service, Whisper allows users to send messages, receive replies using anonymous nicknames. It has grown

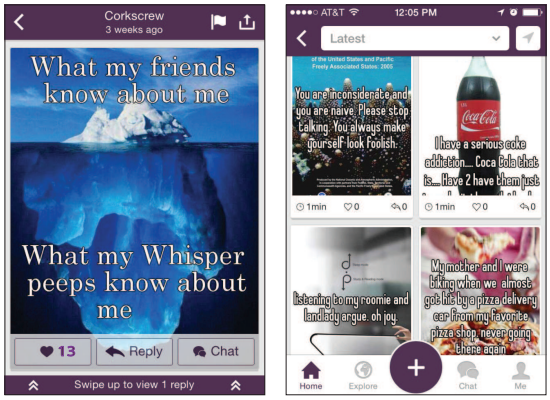


Figure 1: Screenshot of a sample whisper message (left) and the public stream of latest whispers (right).

tremendously in popularity since launching in 2012, and averages more than 3 billion monthly page views as of early 2014 [16]. The functionality is very simple: the app overlays each user's short text message on top of a background image based on keywords from the message (Figure 1). The resulting *whisper* is posted to the public with the user's random or self-chosen nickname. Others can *heart* (Whisper's version of "like") a message anonymously, or post a public followup reply with their own whisper. In addition, users can send *private messages* to the author of a whisper to start a chat, and private messages are only visible to the participants.

User Anonymity. Whisper's focus on anonymity breaks some of the core assumptions made in traditional social networks like Facebook or Google+. First, Whisper users are identified only by randomly assigned (or user-chosen) nicknames, not associated with any personal information, *e.g.*, phone numbers or email addresses[3]. Second, Whisper servers only store public Whispers, and users' private messages are only stored on their end user devices. There is no functionality to search or browse a specific user's historical whispers or replies. Third, there is no notion of a persistent social link between users (*e.g.*, friends on Facebook, followers on Twitter). Thus users are encouraged to interact with a wide range of strangers instead of a known group of "friends."

Public Feeds. Without social links, users browse content from several public lists instead of the news feed of their friends (or followees). These lists include a *latest* list which contains the most recent whispers (system-wise); a *nearby* list which shows whispers posted in nearby areas (about 40 miles of radius range); a *popular* list which only shows top whispers that receive many likes and replies; and *featured* list which shows a subset of popular whispers that are hand-picked by Whisper's content managers. All these lists sort content by most recent first.

2.2 Goals

In its current form, Whisper represents an ideal opportunity to study the impact of pseudo-anonymity on social networks. Three key properties make it ideal for study and analysis. First, Whisper is centralized, *i.e.* there is a single stream of data accessible to all users. Second, Whisper is amenable to periodic data gathering, *i.e.* content is unencrypted and persistent for a moderate amount of time. Third, we were able to meet Whisper's management team,

[3]On the server side, Whisper associates new users with a globally unique identifier (GUID), and binds it to the DeviceID of user's phone. Users can transfer their accounts (private message history) when switching to new phones via iCloud.

and received permission to gather and analyze Whisper's public data stream.

At a high level, our primary goals are to understand how users communicate on pseudo-anonymous social networks, how anonymity affects user behavior, and its consequences on user interactions, long term user engagement, and network stability. Beyond basic analysis of the Whisper network structure, we can solidify our goals into several specific questions. First, how do whisper users interact in an anonymous environment, and do they form communities similar to those in traditional social networks? Second, does Whisper's lack of identities eliminate strong ties between users, and does it eliminate the stickiness critical to long term user engagement in traditional social networks? And given the lack of user-specific network effects, is it possible to model and predict user engagement using short term history cues? Finally, what are the implications of pseudo-anonymity on user content and user privacy?

3. DATA AND INITIAL ANALYSIS

Before diving into our analysis of Whisper, we first describe our data collection methodology and collected datasets. We then describe some high level analyses of our dataset.

3.1 Data Collection

Our goal is to collect whispers and their replies posted in the entire network. Given that Whisper does not archive historical data, our method is to keep crawling newly posted whispers over a long period (February to May 2014). We focus on the "latest" list, which is a public stream of the latest whispers from all Whisper users. Unlike other public lists *e.g.*, "nearby" and "popular", the "latest" list provides access to the entire stream of whispers in the network. Since Whisper does not provide a third-party API, we crawl the "latest" list by scrapping Whisper's website.

Each downloaded whisper includes a whisperID, timestamp, plain text of the whisper, author's GUID, author's nickname, a location tag, and number of received likes and replies. An author's GUID was not intended to act as a persistent ID for each user, but was implemented that way due to Whisper's dependency on a third-party service for private messages. Authors' GUIDs make it possible to track a user's posts over time. After we reported this issue to Whisper's management team, they removed the GUID field in June 2014. The location tag shows user location at the city and state level (*e.g.*, Los Angeles, California), and is available only if the whisper author enabled location sharing permission. Replies to a whisper are similar, the only difference is that replies are also marked with the whisperID of the previous whisper in the thread.

Crawling. We implemented a distributed web crawler with two components, a main crawler that pulls the latest whisper list, and a reply crawler that checks past whispers and collects all sequences of replies associated with an existing whisper. We observe that Whisper servers keep a queue of the latest 10K whispers. Running the main crawler every 30 minutes ensures that we capture all new whispers. In contrast, crawling for replies is more computationally intensive. We crawl for replies every 7 days, and check for new replies for all whispers written in the last month. In practice, we observe that whispers usually receive no followup replies 1 week after being posted.

We ran our crawler from February 6 to May 1, 2014. During this period of roughly 3 months, we collected 9,343,590 total Whispers with 15,268,964 replies and 1,038,364 unique GUIDs. Thanks to server side queues, we collected a continuous data stream despite a small number of interruptions to update crawler code. The only point of note is that, at Whisper's request on April 20, we shifted our crawlers to crawl a different Whisper server using a new set of API calls. The shift reduced load for Whisper, but produced whispers without location tags. Since this only affected 10 days of data, we believe this has little impact on our analysis of location-based features.

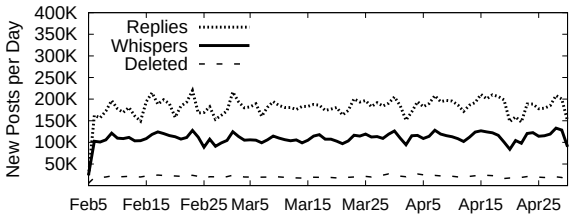


Figure 2: Number of new whispers, new replies and deleted whispers each day.

Validating Consistency. We further verify the completeness of the "latest" stream using a small experiment. We use HTTP requests to simultaneously crawl the "nearby" streams of 6 locations near different cities: Seattle, Houston, Los Angeles, New York, San Francisco and Chicago. We capture these streams for 6 hours, and confirm that the 2000+ whispers from 6 locations were all present in the "latest" stream during the same timeframe.

Limitations. There are two types of data our measurements do not capture. First, we do not capture users who only read/consume whispers but never post any content. Since these passive users do not generate visible user interactions, they are unlikely to affect the majority of our conclusions. Second, our data is limited to visible public data, and we do not have access to private messages between users. Thus our results represent a lower bound on user interactions in the system. As we discuss later, we believe there should be strong correlation between public interactions and private messages.

3.2 Preliminary Analysis

Next we present some high level results on our dataset of whispers, replies and users. Our results in this section set the context for more detailed analysis on user behavior and anonymity in later sections.

Whispers Over Time. We begin by looking at whisper posts over time. Figure 2 shows number of new whispers and replies posted every day during our study. As shown, new content in Whisper is relatively stable, averaging 100K new whispers and 200K replies per day. One interesting observation is that in any time frame, there are significantly more replies than there are original whispers.

During our data collection, we found that a significant portion of whispers is deleted by either the author or Whisper moderators. As far as we can determine, old Whispers do not "expire" and stay on Whisper servers, and can be referenced by following a chain of replies. For deleted whispers, however, we receive an "the whisper does not exist" error when we try to re-crawl their replies. Among the 100K new whispers posted every day, roughly 18% are eventually deleted. We analyze deleted whispers in detail later in §6.

Replies. Users can post replies to a new whisper or other replies. Multiple replies can generate their own replies, thereby forming a tree structure with the original whisper as the root. Figure 3 and Figure 4 show total replies per whisper and the longest chain length (maximum tree depth) per whisper. Unsurprisingly,

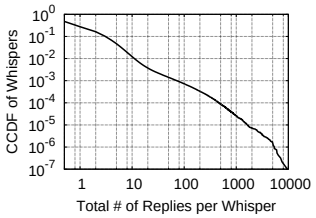


Figure 3: Total number of replies per whisper.

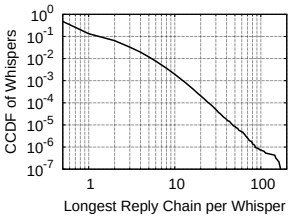


Figure 4: Length of longest reply chain per whisper.

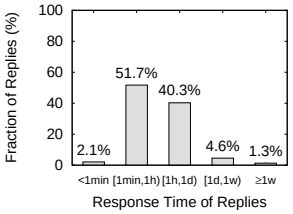


Figure 5: Time gap between reply and original whisper.

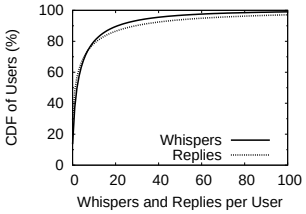


Figure 6: Whispers and replies posted per user.

55% of whispers receive no replies. Since all whispers are posted to the same public lists, each whisper only has a short time window to attract users' attention. Among whispers with replies, roughly 25% have a chain of at least 2 replies. These essentially become threads of conversations between users.

Figure 5 plots the distribution of reply arrival time, which is the time gap between each reply and the original whisper. 54% of replies arrive within an hour of the original whisper, and more than 94% of replies arrive within a day. Only 1.3% of replies arrive a week or more after the whisper. This confirms our intuition—if a whisper does not get attention shortly after posting, it is unlikely to get attention later.

Users. We look at content-generated per user based on unique GUID. Figure 6 plots the number of whispers and replies posted by each user. Most users (80%) post less than 10 total whispers or replies. Roughly 15% of users only post replies but no original whispers, and 30% of users only post whispers but no replies.

Content Analysis. A high-level analysis of the contents of whispers shows that users post highly personal content. A search of singular first-person pronouns (*e.g.*, I, me, my, myself) hits about 62% of all whispers. We also find a heavy usage of emotional key words. Specifically, 40% of whispers contain one of the 1,113 human mood related key words provided by WordNet Affect [33]. Finally, people often ask questions seeking advice or empathy. About 20% of whispers are questions, based on the usage of question marks and interrogatives (*e.g.*, what, why, which). These three categories effectively cover 85% of all whispers. It is clear that the anonymity provided by Whisper encourages users to post personal and intimate content without privacy concerns. We will take a closer look at "topics" of whispers in §6.

4. USER INTERACTIONS

Our study begins with user interactions on Whisper. The fact that Whisper users cannot construct persistent social links between them fundamentally changes how users interact and develop friendships. In this section, we study the bidirectional interactions built from whispers and their replies, and seek to understand user interactions from three different levels. First, we study interactions at a global network level, by comparing structural properties of the Whisper interaction graph to those of traditional OSNs, *e.g.*, Facebook and Twitter. Second, we study user communities in the Whisper graph and explore key factors driving their formation. Finally, we look at interactions at per-user level to understand if users still develop strong ties (frequently interacted friends) in Whisper.

4.1 Whisper Interaction Graph

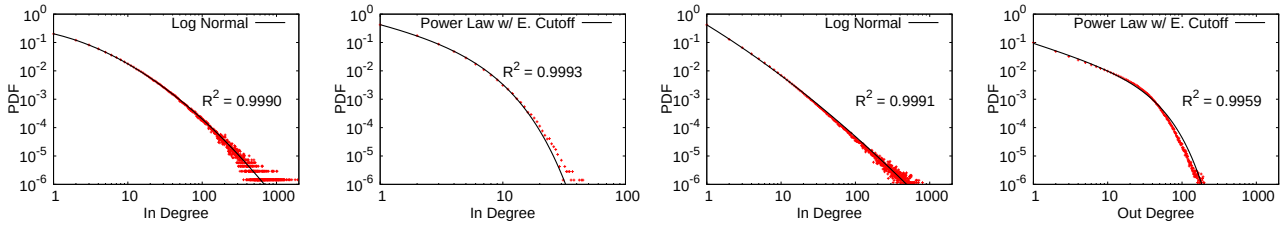
We first compare the interaction graph of Whisper with those of traditional online social networks (Facebook and Twitter). Our goal is to understand whether the lack of social links in Whisper fundamentally changes users' interaction patterns at an aggregate network level. We build a Whisper interaction graph based on whispers and replies, and compare its structure to those of graphs constructed from Facebook wall posts and Twitter retweets.

Building Interaction Graphs. We build the Whisper interaction graph based on whispers and followup replies, which are the primary publicly visible interactions in Whisper. The result is a directed interaction graph, where nodes are users and edges represent reply actions. For example, if user A posts a reply whisper to B's whisper, we build a directed edge from A to B. Only direct replies are used to build edges. We remove disconnected singleton nodes from the graph. We produce a main interaction graph from our 3 month dataset (*Whisper-all*).

For comparison, we also build interaction graphs for Facebook and Twitter, based on anonymous datasets from our prior work [39, 42]. Both datasets crawled historical data that covers user interactions over at least 3 months. We built a directed interaction graph using Facebook wall post data: if user A posts on user B's wall, we create a directed edge from A to B. For Twitter, we built the graph based on retweet interactions: if user A retweets a tweet from B, we create a directed edge from A to B. To match the 3-month time coverage of Whisper graph, we build similar Facebook and Twitter graphs each using data covering 3 month periods. Table 1 shows the key statistics of all three interaction graphs.

Degree Distribution and Fittings. Users in Whisper show much higher average degree than users in Facebook and Twitter, meaning users interact with a larger sample of other users. We determine the best fitting function for each graph's degree distribution using 3 commonly used fitting functions for social graphs, power law ($P(k) \propto k^{-\alpha}$), power law with exponential cutoff ($P(k) \propto k^{-\alpha}e^{-\lambda k}$) and lognormal ($P(k) \propto e^{\frac{(lnx-\mu)^2}{2\sigma^2}}$) [14, 39]. We follow the fitting method in [10] and use Matlab to compute fitting parameters and accuracy (R-squared values), and show the results in Figure 7. For both the Whisper and Facebook graphs, the out-degree distribution looks similar to the in-degree distribution. For brevity, we only show the in-degree distribution for each graph. Intuitively, Facebook was designed to emulate offline social relationships, and the prevalent bidirectional interactions lead to symmetric in- and out-degree distributions. For Whisper, user interactions are largely random between users. In contrast, Twitter's in-degree and out-degree distributions are significantly different. It's well known that Twitter is more of an information dissemination medium than a social network, and interactions are highly asymmetric [25].

Clustering Coefficient. Clustering coefficient is the ratio of the number of connections that exist between a node's immediate neighbors over all possible connections that could exist. It measures the level of local connectivity between nodes. Clustering coefficient in the Whisper graph (0.033) is much smaller than that of Facebook (0.059) and Twitter (0.048). The cause is clear: Whisper



(a) Whisper In (b) Facebook In (c) Twitter In (d) Twitter Out

Figure 7: Degree distribution and fitting result.

Graph	# of Nodes	# of Edges	Avg. Degree	Clustering Coef.	Avg. Path Length	Assortativity Coef.	Largest SCC	Largest WCC
Whisper	690K	6,531K	9.47	0.033	4.28	-0.011	63.3%	98.9%
Facebook	707K	1,260K	1.78	0.059	10.13	0.116	21.2%	84.8%
Twitter	4,317K	16,972K	3.93	0.048	5.52	-0.025	14.2%	97.2%

Table 1: High level statistics of different interaction graphs.

users are highly likely to interact with complete strangers, who are highly unlikely to interact with each other.

Average Path Length. Average path length is the average of all pairs of shortest paths in the graph. Given the size of our graphs, it's impractical to compute the shortest path for all node pairs. Instead, we randomly select 1000 nodes in each graph and compute the average shortest path from them to all other nodes in the graph. The result shows that Whisper graph has the shortest average path length of the 3 networks. This is again intuitive, since the formation of interactions between random strangers creates numerous shortcuts in the graph, thereby shrinking the average path length. Considering Whisper's high average degree, low clustering level and short average path length, Whisper exhibits more properties of a random graph [38] than those of a "small-world" network like Facebook and Twitter.

Assortativity. Assortativity coefficient measures the probability for nodes in a graph to link to other nodes of similar degrees. Assortativity > 0 indicates that nodes tend to connect with other nodes of similar degree, while assortativity < 0 indicates that nodes connect to others with dissimilar degrees. Our result shows the assortativity coefficient of Whisper graph is very close to zero (-0.011), which closely resembles a random graph [29]. In contrast, similar users tend to flock together in social networks with bidirectional links (*e.g.*, Facebook), producing positive assortativity (0.116). In Twitter, large numbers of normal users follow celebrities and notable figures, thus producing a more negative assortativity (-0.025).

4.2 Communities in the Interaction Graph

Next, we analyze the presence of community structures in Whisper's interaction graph. Communities are defined as groups of nodes that are densely connected within but sparsely connected to the rest of the network, *i.e.* high modularity. We seek to answer two key questions. First, without persistent social links, do Whisper users still form communities in the interaction graph? Second, if so, what's the key factor driving the formation of user communities?

Community Detection. We start by applying community detection algorithms to Whisper graphs to examine whether community structures exist. We choose two widely used community detection algorithms Louvain [7] and Wakita [37]. We compute the *modularity* of the resulting communities. Modularity [28] is a well accepted metric used for community detection, which measures the difference between the fraction of links within the communities and

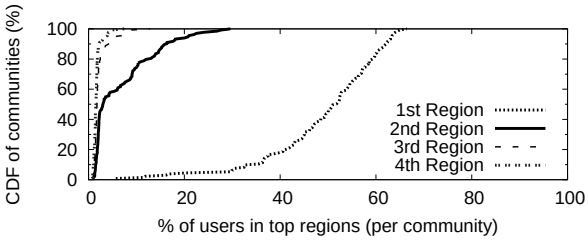


Figure 8: % of users in top regions per community. Users within a community are highly skewed to one region.

Community (size)	Top 4 Region (% of users)
C_1 (61,686)	NY (11), NJ (10), CT (4.8), CA (4.2)
C_2 (39,824)	England (61), Wales (3.5), CA (1.1), TX (0.9)
C_3 (28,342)	CA (62), TX(1.5), England (1.2), AZ (0.9)
C_4 (22,010)	IL (37), WI (21), IN (4.5), CA (1.5)
C_5 (16,017)	CA (64), England (1.4), TX (1.3), NY (0.8)

Table 2: Top 5 biggest communities and their top regions.

the expected fraction when links are randomly connected. Modularity ranges from -1 to 1, and higher values represent stronger communities in the graph.

To capture user interactions using the graph, we weigh graph edges based on the number of interactions between the two nodes. Also we focus this analysis on the biggest weakly connected component, which contains 99% of all nodes. Applying Louvain produces average modularity of communities of 0.4902 for Whisper. In practice, modularity > 0.3 indicates significant community structure in a graph [24]. We confirm our results using the Wakita community detection algorithm, and find a resulting modularity of 0.409 (also above 0.3). As a point of reference, modularity scores of existing social graphs include Facebook (0.63), Youtube (0.66) and Orkut (0.67) [24]. Not surprisingly, the relatively weak communities in Whisper match other observations including low clustering activities and weak ties.

User Communities vs. Geolocation. The natural followup question is, why are there any communities in Whisper at all? If user interactions are random, then shouldn't all interactions be uniform? Our hypothesis is that this is due to the "nearby" functionality in Whisper, which allows users to browse (and likely reply to) whispers posted by people in nearby areas. Our intuition is that the nearby stream drives users to interact more often with others

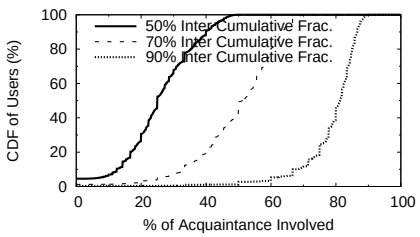


Figure 9: The distribution of users' interaction among their acquaintances, for different % of interactions.

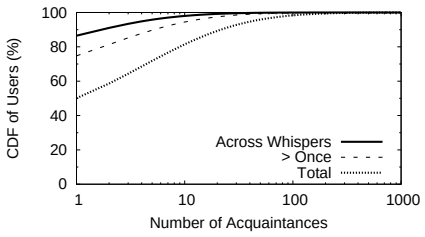


Figure 10: Number of user's acquaintances, and those that users interact > once and across whispers.

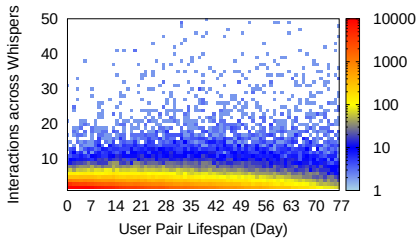


Figure 11: Users pairs with interaction across whispers: lifespan vs. # of interactions.

in the same geographic locale, thus helping geographically driven communities form in the interaction graph.

To test this idea, we examine the most prevalent geographic regions for users in each community. If geo-factor is the key driving force to form communities, then the community should be dominated by users from the same location. Table 2 shows the top 5 Whisper communities generated by Louvain and their corresponding top regions. Here we use "state" or "province" level location tags. We find that the top 5 communities all confirm this: the majority of users are skewed to a single region or several geographically adjacent regions (*e.g.* NY, NJ and CT for C_1).

To quantify this phenomenon across all communities, we plot in Figure 8 the fraction of users in the top four geographic regions per community. Louvain produces 912 communities with varying sizes, and we only consider the largest 150 communities which together cover >90% of users. Again the results confirm our hypothesis, community membership is dominated by the top region or top 2 regions. This strong geographic locality in interactions confirms that the "nearby" stream plays an important role in the formation of Whisper communities. While other factors may contribute to the formation of user communities (*e.g.*, users' shared topics and interests, time zones), we leave their analysis to future work.

4.3 User Interactions and Strong Ties

Finally, we analyze user interactions and implicit social links at the per-user level. Recall that Whisper's lack of persistent identities and social links encourages users to interact with strangers. In the following, we seek the answers to two key questions. First, do users have a fixed set of "friends" that they frequently interact with? Such friendships could have formed despite the anonymous nature of Whisper nicknames. Second, how likely are any strong ties the result of offline friendships?

Per-user Interaction. We search for potential friendships (*i.e.* strong ties) by looking for pairs of users who interact more frequently with each other than with others. For convenience, we call the set of people that a user interacts with (regardless of direction) as her *acquaintances*. For each user, we compute a distribution of her interactions across her top acquaintances, and look for *skew* in her interactions with all acquaintances. We select several points (50-, 70- and 90-percentiles) from each user's distribution and aggregate them in a CDF to show the percentage of top acquaintances involved (Figure 9). To avoid statistical outliers, we only include users with at least 10 interactions.

We find user's interactions are distributed rather evenly across acquaintances. Take the 90-percentile line for example, for nearly all the users (∼90%), more than 70% of their acquaintances are responsible for 90% of their interactions. This relatively low skew in Whisper is exactly the opposite of traditional OSNs like Facebook, where a small fraction of friends (strong ties) are responsible for the vast majority of user's interactions [39].

Interaction across Whispers. Across a user's acquaintances, we look for potential strong ties, *i.e.* acquaintances with whom the user interacts often. Figure 10 shows user's number of total acquaintances, acquaintances that users interacted more than once, and acquaintances that users interact more than once *using multiple whisper threads*. In Whisper, it's common for people to interact more than once under the same whisper. However, it's rare to talk with the same person across different whispers, because keeping track of a particular user via their anonymous nickname is difficult. As shown in Figure 10, only 13% of users have acquaintances that they interact with across whispers.

We then select those user pairs who have interacted across whispers for further analysis. In total, there are 503K such user-pairs. Figure 11 presents the heat map of these user-pairs' lifespan (timespan between their first and last interaction) and their number of interactions across whispers. Note that the color palette is log-scale—the vast majority of user pairs are stacked at the left bottom corner, indicating short-lived, low-interaction relationships. Only a very small fraction of outliers (right top corner) achieved long-term and frequent interactions.

Friends or Random Encounters? Even though the strong ties are outliers, it is interesting to explore how could these user-pairs constantly interact with each other *across whispers*: Are these pairs of offline friends who actively track each other in the public feeds (using nicknames), or are these simply users who bump into each other often by chance? We realize this is a very hard question to answer deterministically. But we have a key intuition: if these interactions are truly random, then it is highly likely that these two users are co-located in same geographic area, particularly areas with a sparse population of Whisper users. Then as long as the two users actively post whispers, they have a good chance to see each other in the nearby list.

Now we use our data to test this intuition. For user-pairs with cross-whisper interactions, we first examine their geographic distances[4]. We find that among 503K user pairs, 90% have two users co-located in the same "state" and 75% have their distance <40 miles which is the maximum range of the nearby stream. Figure 12 shows the correlation between geo-distance and the interaction frequency of user pairs. Each stacked bar adds up to 100%, and each category represents user pairs with different interaction level (*i.e.* number of interactions across whispers). It shows that frequent interactions are more skewed to users that are geographically close to each other.

[4]This is the distance between two user's city-level tags. The GPS coordinates of each city are obtained from Google Geocoding API.

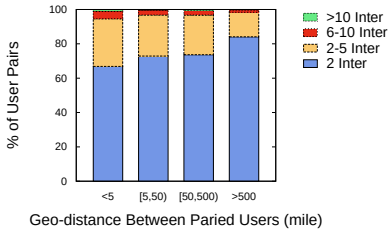


Figure 12: For all user pairs, distance between two users vs. # of interactions of the user pair.

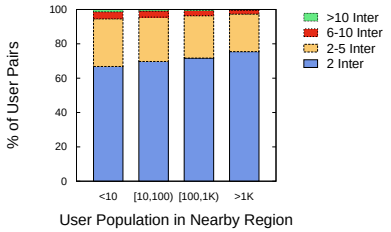


Figure 13: For nearby user pairs, user population in nearby areas vs. # of interaction of the user pair.

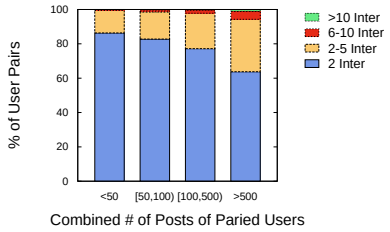


Figure 14: For nearby user pairs, total # of whispers vs. # of interactions of the user pair.

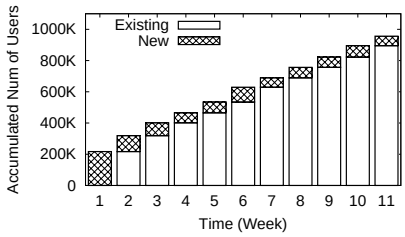


Figure 15: The growth of user population in our dataset over time.

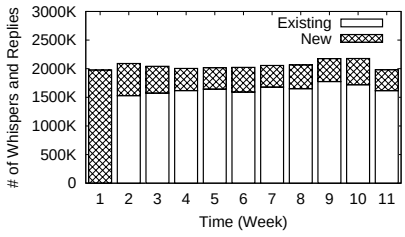


Figure 16: # of whispers and replies by new and old users per week.

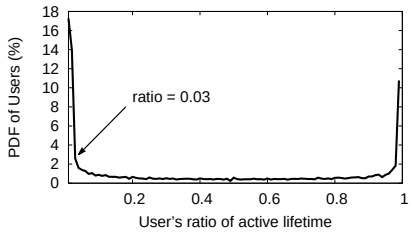


Figure 17: User's active lifetime over staying time in our dataset.

Then we further examine these pairs co-located in nearby areas (*i.e.* distance <40 miles). More specifically, we analyze two factors that potentially impact users' likelihood of chance encounters—the user population in the geographic area and total number of whispers posted by the two users (Figure 13 and Figure 14). Intuitively, the smaller user population in the same nearby area, the higher chance to encounter the same person in the nearby list again and again. Similarly, the more whispers two users post, the more likely they encounter each other and form interactions. Here the user population is estimated by the total number of unique users that have the same city-level location tags with the paired users. Both results confirm our intuition. As user population density decreases and as the number of user posts increases, the probability of more frequent user-pair interactions also increases.

In summary, our analysis suggests strong ties are extremely rare in Whisper. We also find strong ties are skewed to user-pairs who have a higher chance to encounter each other (*i.e.* active users that are co-located in areas with sparse user population). Thus while it is possible to develop strong relationships from Whisper interactions, such relationships are likely heavily influenced by geographic density and user whisper frequency. Note that our analysis relies on only public interactions and do not include private messages. Intuitively, we believe users' private interactions should correlate with their public interactions, and we can predict user pairs with private interactions from their public interactions. Prior work also confirms that public interactions are more informative when modeling strength of ties than private communications [13, 22].

5. USER ENGAGEMENT

Thus far our analysis shows Whisper users tend to interact with strangers rather than stable friends. The negative consequence is that a lack of strong ties usually produces a less "sticky" network, *i.e.* fewer disincentives to prevent users from leaving [11]. This raises a natural question: without strong ties, can Whisper users stay engaged in the network in the long run?

In this section, we seek to consider this question by looking at *per-user* engagement. First, we examine user engagement over time to understand user attrition in the 3 month period of our dataset. Second, we evaluate a machine learning classifier and show that we can accurately predict whether users stay engaged in the system using only a short history of their actions after their first post. We use experiments to determine key signals that indicate a user's intention to leave. Note that our analysis is limited to "active" users who have posted at least 1 whisper or reply, and does not include passive users who consume but do not contribute content.

5.1 User Engagement Over Time

We start with basic analysis of user activity over time using three metrics: user population growth, content contribution by new versus existing users, and the distribution of users' active lifetime.

User Population Growth. Figure 15 shows the total number of users over time (11 weeks) in our dataset. Each bar shows a breakdown of new users who just joined that week (new) and the existing users we observed before that week (existing). We observe a stable arrival rate of new users to the network, roughly 80K new users per week. Recall that the daily new posts (whispers and replies) in the entire network remain roughly stable (see Figure 2), despite the growth in users. This indicates there are an ongoing number of users who "disengage," *i.e.* stop posting whispers or replies.

Content by New and Existing Users. This motivates us to look at the relative contribution of content by new and existing users. Figure 16 shows the breakdown of posts (whispers plus replies) by users who showed up for the first time in the current week (new) and users who showed up before this week (existing). We find that new users make significant contributions to the overall whisper stream ($> 20\%$). However, as more and more users transition from new to "existing users," content generation by existing users does not grow significantly over time. This confirms our intuition that a certain portion of users are disengaging over time.

Per-user Active Period. Next, we focus on individual users and examine how long users stay active before they disengage. More specifically, we compute their active "lifetime" (timespan between their first and last posts) over their staying time in the dataset (timespan between a user's first post and the last date of our data collection). Given our focus on long-term activity, we exclude users who just recently joined during the last the month of our data collection. Thus for Figure 17, we only consider users who have been in our dataset for at least one month (70.3% of all users).

Figure 17 shows the distribution of user's ratio of active lifetime (PDF). Users are clearly clustered into two extremes: one major cluster around an extremely low ratio (0.03), representing those who quickly turned inactive in 1 or 2 days after their first post; another major cluster around 1.00, representing users who remain active for their entire time in the dataset (at least 1 month). Similar patterns have also been observed in other user generated content (UGC) networks, such as blogs and Q&A services [17]. If we set a threshold for active ratio at 0.03, these "try and leave" users account for 30% of all users. This explains our observation in Figure 16—because a significant portion of users become inactive quickly, the overall content posting rate remains stable despite a significant number of new users joining the network.

5.2 Predicting User Engagement

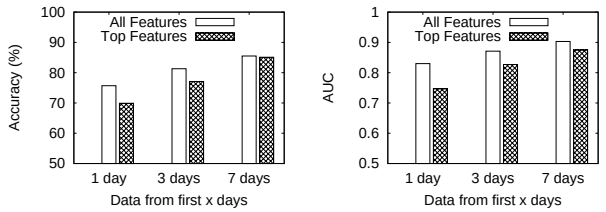
A key observation of the above analysis is that Whisper users tend to fall into one of two behavioral extremes—either staying active for a long time, or quickly turning inactive (Figure 17). The bimodal nature of the distribution hints at the potential to classify users into the two clusters.

Here, we experiment with machine learning (ML) classifiers to determine if we can predict long term user engagement based on their early behavior after their first post (in our dataset). We seek to answer three key questions: First, is this prediction even possible? Second, what ML models produce the most accurate predictions? Third, what early-day signals can most strongly indicate a user's intention to leave?

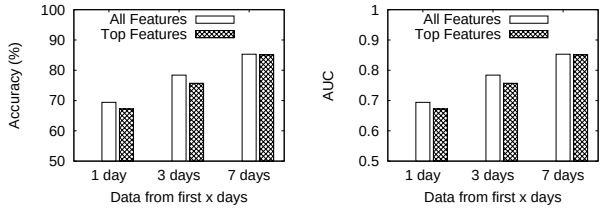
We take three steps to answer the above questions. First, we collect a set of behavioral features based on users' activities in their first X days on Whisper, ideally with a small value for X. Second, we use these features to build different machine learning classifiers to predict long term user engagement. Finally, we run feature selection to determine the features that provide the best early signals indicating which users might disengage.

Features. We explore multiple different classes of features (20 features in all) to profile users' behavior during their first X days. Out of these, we will select the most essential features.

- *Content posting features (F1-F7).* 7 features: user's number of total posts, number of whispers, number of replies, number of deleted whispers, and number of days with at least one post/whisper/reply.
- *Interaction features (F8-F15).* 8 features: ratio of replies in total posts, number of acquaintances, number of bi-directional acquaintances, outgoing replies over all replies, maximum number of interactions with the same user, ratio of whispers with replies, and average number of replies and likes per whisper.
- *Temporal features (F16-F17):* 2 features: average delay before first reply to user's whisper; average delay of user's replies to other users' whispers.
- *Activity trend (F18-F20):* 3 features: we equally split each user's first X days into three buckets and record the number of posts in each bucket ($First$, $Middle$ and $Last$). We compute 2 features as $\frac{Middle}{First}$ and $\frac{Last}{First}$. Finally, whether the number of posts decreases monotonically across the three buckets.



(a) Predicting Inactive vs. Active (RF)



(b) Predicting Inactive vs. Active (SVM)

Figure 18: Prediction result using Random Forests and SVM. The model performance is evaluated by accuracy (left) and Area under ROC curve (right) .

Classifier Experiments. To build a training set for our classifiers, we focus on users that have at least a month's worth of activity history in our dataset (730K users). We select a set of "short-term" users who tried the app for 1-2 days and quickly disengaged (no more posts). Using results from Figure 17, we randomly sample 50K users from those whose active lifetime ratio < 0.03 as the *Inactive* set. We then choose a random sample of 50K users whose active lifetime ratio > 0.03 to form the *Active* set.

Our goal is to classify the two sets of users solely based on users' activities in their first X days, and we use 1, 3 and 7 as values of X. We build multiple machine learning classifiers including Random Forests (RF), Support Vector Machine (SVM) and Bayes Network (BN), using implementations of these algorithms in WEKA [19] with default parameters. For each experiment, we run 10-fold cross validation and report classification accuracy and area under ROC curve (AUC). Accuracy refers to the ratio of correctly predicted instances over all instances. AUC is another widely used metric, with higher AUC indicating stronger prediction power. For instance, AUC > 0.5 means the prediction is better than random guessing.

The experiment results with Random Forests and SVM are shown in Figure 18. The Bayesian results closely match those of SVM, thus we omit them for brevity. We make two key observations. First, behavioral features are effective in predicting future engagement. The accuracy is high (75%) even when only using users' first-day data (RF). This confirms that users' early actions can act as indicators of their future activity. If we include a week's worth of data, we can achieve accuracy up to 85%. Second, we find different classifiers achieve similar performance given 7 days of data. However, their results diverge when they are constrained to using less data (*e.g.*, 1-day). With less data, Random Forests produce more accurate predictions than SVM and Bayesian networks.

Feature Selection. Finally, we seek to identify the most powerful signals to predict a user's long-term engagement. To find the answer, we perform feature selection on the 20 features. More specifi-

Rank	Observation Time Frame		
	1 day	3 days	7 days
1	Interact-F9 (0.15)	Post-F5 (0.27)	Post-F5 (0.46)
2	Interact-F11 (0.12)	Trend-F19 (0.18)	Post-F6 (0.31)
3	Interact-F10 (0.11)	Post-F6 (0.18)	Trend-F19 (0.28)
4	Interact-F12 (0.11)	Interact-F9 (0.16)	Post-F1 (0.27)
5	Trend-F18 (0.05)	Post-F1 (0.16)	Post-F7 (0.23)
6	Interact-F15 (0.04)	Post-F7 (0.13)	Trend-F20 (0.21)
7	Post-F1 (0.04)	Interact-F15 (0.12)	Interact-F15 (0.21)
8	Interact-F8 (0.04)	Interact-F11 (0.12)	Post-F2 (0.19)

Table 3: The top 8 feature and its categories ranked by information gain (values shown in parentheses).

cally, we rank features based on *Information Gain* [18], which measures feature's distinguishing power over the two classes of data. We list the top 8 features in Table 3. As expected, prediction power varies significantly, and information gain drops off quickly (particularly for 1 day) after the top 4 features. To validate their prediction power, we repeat each experiment with only their top 4 features. The results in Figure 18 show that the top 4 features achieve most of the accuracy of the entire classifier, but with much less complexity.

Then we take a closer look at the top features. First, we note that the 1-day classifier relies on different set of features compared with 3- and 7-day classifiers. The 1-day models rely heavily on *interaction features*. Intuitively, the model predicts whether a user will stay engaged based on how actively the user participates in social interactions. If a user received many replies or actively replied to others on her first day, there's a high chance for this user to stay longer. For 3- and 7-day models, we find that the key features shift to user's *content posting* and *activity trend* features. This means once we monitor the users for a longer period, the user's intention to stay or leave can be more accurately reflected in her posting frequency and volume, and whether that activity is declining over time.

Engaging Users with Notifications. Stimulating user engagement is a key goal for any new service. One tool Whisper has already deployed is push notifications that deliver the "whisper of the day" to users' mobile device every evening between 7 and 9pm. The exact notification time varies each day and between Android and iOS devices. To examine the impact of these notifications, we conduct a small experiment. We monitor the notification time on 5 different phones every day for 6 days. We look at user activity in the Whisper stream for 5 minute and 10 minute intervals following the notifications, and find no statistically significant increase in new replies or whispers compared to other 5 or 10 minute windows between 7 and 9pm. This means that while these notifications may serve to engage users to read popular whispers, there is no significant increase in new whispers or replies as a result.

6. CONTENT MODERATION IN WHISPER

Anonymity facilitates free speech, but also inevitably fosters abusive content and behavior [21, 35]. Like other anonymous communities, Whisper faces the same challenge of dealing with abusive content (*e.g.*, nudity, pornography or obscenity) in their network. In addition to a crowdsourcing-based user reporting mechanism, Whisper also has dedicated employees to moderate whispers [16]. Our basic measurements (§3.2) also suggest this has a significant impact on the system, as we observed a large volume of whispers ($>$1.7 million) has been deleted during the 3 months of our study. The ratio of Whisper's deleted content (18%) is much higher than traditional social networks like Twitter ($<$4%) [1, 30].

Topic	**Top 50 Keywords Most Related to Deleted Whispers**
Sexting (36)	sext, wood, naughty, kinky, sexting, bj, threesome, dirty, role, fwb, panties, vibrator, bi, inches, lesbians, hookup, hairy, nipples, freaky, boobs, fantasy, fantasies, dare, trade, oral, takers, sugar, strings, experiment, curious, daddy, eaten, tease, entertain, athletic
Selfie (7)	rate, selfie, selfies, send, inbox, sends, pic
Chat (7)	f, dm, pm, chat, ladys, message, m
Topic	**Top 50 Keywords Least Related to Deleted Whispers**
Emotion (17)	panic, emotions, argument, meds, hardest, fear, tears, sober, frozen, argue, failure, unfortunately, understands, anxiety, understood, aware, strength
Religion (10)	beliefs, path, faith, christians, atheist, bible, create, religion, praying, helped
Entertain. (8)	episode, series, season, anime, books, knowledge, restaurant, character
Life story (6)	memories, moments, escape, raised, thank, thanks
Work (5)	interview, ability, genius, research, process
Politics (1)	government
Others (3)	exactly, beginning, example

Table 4: Topics of top and bottom 50 keywords related to whisper deletion.

In this section, we take a closer look at content deletions in Whisper. First, we analyze the content of deleted whispers to infer the reasons behind deletions. Second, we analyze the lifetime of deleted whispers to understand how fast do whispers get deleted. Third, we focus on authors of deleted whispers and compare their behavior to the norm.

Before we begin, we note that while users can delete their own whispers, we believe server-side content moderation is responsible for the large majority of missing whispers in our data. Intuitively, users who reconsider and later delete their own whispers are likely to do so within a relatively short time frame. In contrast, our "deleted" dataset comes from our followup crawl for replies, which runs once a week. In fact, since our main crawler on the latest stream runs every 30 minutes, we expect most self-deleted whispers will not even show up in our core dataset.

Content Analysis of Deleted Whispers. To explore the reasons behind deletion, we analyze the content of deleted whispers. Since whispers are usually very short, Natural Language Processing (NLP) tools do not work well (we confirmed via experiments). Thus we take a keyword-based approach: we extract keywords from all whispers and examine which keywords correlate with deleted whispers. First, before processing, we exclude common stopwords[5] from our keyword list. Also to avoid statistical outliers, we exclude low frequency words that appear in less than 0.05% of whispers. Then for each keyword, we compute a *deletion ratio* as the number of deleted whispers with this keyword over all whispers with this keyword. We rank keywords by deletion ratio, and examine the top and bottom keywords.

We run this analysis on all 9 million original (not including replies) whispers in our dataset, 1.7M of which are later deleted. This produces 2324 keywords ranked by deletion ratio. We list the top and bottom 50 keywords in Table 4 and classify them manually into topic categories. Not surprisingly, many deleted whispers violate Whisper's stated user policies on sexually explicit messages and nudity. In contrast, topics related to personal expression, religion, and politics are least likely to be deleted.

[5] `http://norm.al/2009/04/14/list-of-english-stop-words`

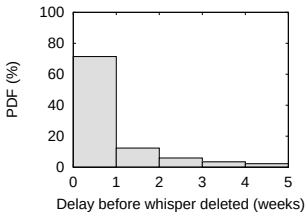


Figure 19: Deletion speed (coarse-grained).

Figure 20: Deletion speed (fine-grained).

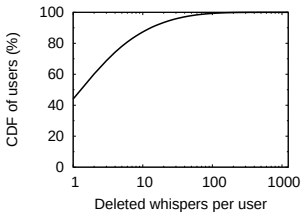


Figure 21: # of Deleted whispers per user.

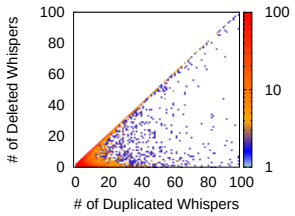


Figure 22: Duplicated vs. deleted whispers.

Deletion Delay. Next we analyze the deletion delay of whispers, *i.e.* how long do whispers stay in the system before they are deleted? Recall that our reply crawler works once a week, and thus detects deleted whispers on the granularity of once a week. As shown in Figure 19, the majority (70%) of deleted whispers are "deleted" within one week after posting. A small portion (2%) of whispers have stayed for more than a month before deletion. Since most whispers lose user attention after one week (Figure 5), we believe these deletions are not the results of crowdsourcing flagging, but deleted by Whisper moderators.

To get a more fine grain view of whisper deletions, we perform a period of frequent crawls on a small set of whispers. On April 14, 2014, we select 200K new whispers from our crawl of the latest whisper stream, and check on (recrawl) these whispers every 3 hours over a period of 7 days. Of the 200K whispers, 32,153 whispers are deleted during our monitoring period (a week). The more fine-grained distribution of the lifetime (hourly) of these whispers is shown in Figure 20. We find the peak of whisper deletion to be between 3 and 9 hours after posting, and the vast majority of deletions happen within 24 hours of posting. This suggests that the moderation system in Whisper works quickly to flag and remove offensive whispers. However, it is unclear whether this level of responsiveness is sufficient, since user page views focus on the most recent whispers, and moderation after 3 hours is possibly too late to impact the content most users see.

Characterizing Authors of Deleted Whispers. Finally, we take a closer look at the authors of deleted whispers to check for signs of suspicious behavior. In total, 263K users (25.4%) out of all users in our dataset have at least one deleted whisper. The distribution of deleted whispers is highly skewed across these users: 24% of users are responsible for 80% of all deleted whispers. The worst offender is a user who had 1230 whisper deleted during the time period of our study, while roughly half of the users only have a single deletion (Figure 21).

We observed anecdotal evidence of duplicate whispers in the set of deleted whispers. We find that frequently reposted duplicate whispers are highly likely to be deleted. Among our 263K users with at least 1 deleted whisper, we find 25K users have posted duplicate whispers. In Figure 22, we plot each user's number of duplicated whispers versus the number of deleted whispers. We observe a clear clustering of users around the straight line of $y = x$. This indicates that when users post many duplicated whispers, there's a higher chance that most or all duplicated whispers are deleted.

We also observe that authors of deleted whispers change their nicknames more often than the average user. Figure 23 shows the distribution of total number of nicknames used by each user. We categorize users based on how many deletions they have, and also include a baseline of users with 0 deletions. We find users with no deletion rarely change their nicknames, if ever, but nickname changes occur far more frequently for users with many deleted whispers. We speculate that perhaps users change their nickname to avoid being flagged or blacklisted. Since users cannot see their own GUID when using the app, they may assume the system identifies them using only their nickname.

7. TRACKING WHISPER USERS

In the final component of our Whisper study, we take a close look at a vulnerability that exposes detailed location of Whisper authors to the system. In practical terms, this attack allows a Whisper user to accurately track (or potential stalk) another Whisper user through whispers they've written, by writing simple scripts that query Whisper servers. This attack demonstrates the inherent risks to user privacy in mobile applications, even for apps that target user anonymity as a core goal. Note that we met the Whisper team in person and informed them of this attack. They are supportive of this work, and have already taken steps to remove this vulnerability.

In this section, we describe details of this location tracking attack. The attack makes use of Whisper's "nearby" function, which returns a list of whispers posted nearby, attaching a "distance" field to each whisper. The attack generates numerous "nearby" queries from different vantage points, and uses statistical analysis to reverse engineer the whisper author's location. We validate the efficacy of this attack through real-world experiments.

7.1 Pinpointing User Locations

We start by describing the high-levels of the attack: when a user (*i.e.* the victim) posts a new whisper, he exposes his location to the Whisper server. An attacker in an nearby area can query the nearby list to get their "distance" to the whisper author. The methodology is simple: the attacker can move to different (nearby) locations and query the nearby list for the distance to the victim. Using multiple distance measurements, the attacker can *triangulate* the whisper author's location. The fact that Whisper does not authenticate location in its queries makes this easier, an attacker can issue numerous distance queries from different locations all while sitting in the comfort of her living room.

With a bit more effort, an attacker can even track the victim's movement over time, by triangulating his location every time he posts a whisper. In practice, this means the attacker can physically go and stalk the victim. While the effective error is roughly 0.2 miles (details below), it is more than sufficient to infer the victim's movement to specific points of interest. Considering most Whisper users are young adults or teenagers [4], this attack can lead to severe consequences.

Distance Granularity and Errors. Implementing this attack is nontrivial. Whisper's design team has always been aware of location tracking risks to its users, and built in basic defense mechanisms into the current system. First, they apply a distance offset to every whisper, so the location stored on their servers is always

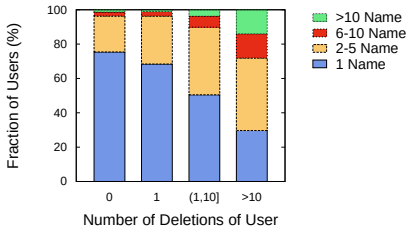


Figure 23: User's number of deletions vs. number of nicknames.

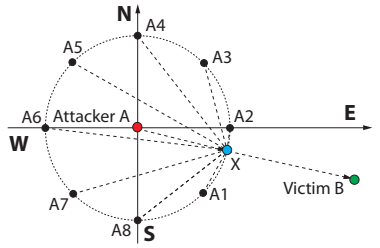


Figure 24: Estimating the distance and direction to the victim.

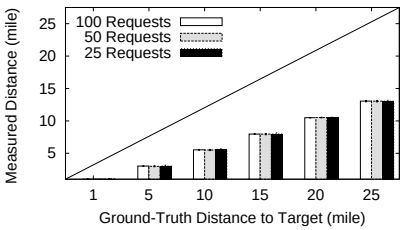


Figure 25: True distance vs. measured average distance (>1 mile).

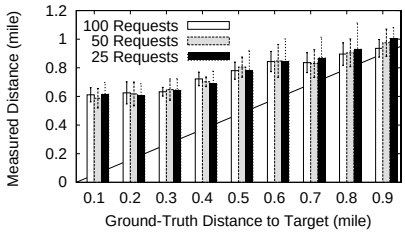


Figure 26: True distance vs. measured average distance (within 1 mile).

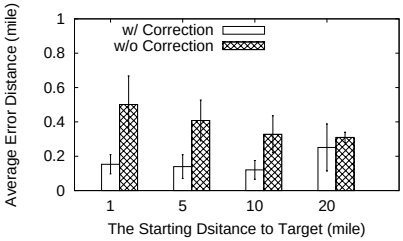


Figure 27: The final error distance of the attack.

Figure 28: Number of hops to approach the victim.

off by some distance to the actual author location. Second, the distance field returned by the nearby function is a coarse-grained integer value (in miles). This was a recent change made by Whisper in February 2014, before which the nearby function returned distances with decimal values. Third, Whisper server adds a random error to the answer to each query, *i.e.* when we query the nearby list repetitively from the same location, each query returns a different distance for the same whisper. The specific error function is unknown.

Attack Details. To accurately pinpoint a user location, our approach is to extensively measure the "distance" from different vantage points, and use large-scale statistics to infer user's location. Specifically, our attack exploits a key property of Whisper: servers allow anyone to query the nearby list with arbitrarily self-reported GPS values as input, and impose no rate limits on such queries. This effectively helps us to overcome the limitations (*i.e.* random error, coarse granularity) on the returned distance. First, we can reduce or eliminate per-query noise by taking the average distance across numerous queries from the same observation location. Second, even though the absolute distance is still not accurate, we can estimate the *direction* to the victim based on the measurements from different locations. Then with distance and direction, an attacker can repeat the measurement from a location closer to the victim, thus iteratively deducing the victim's real location.

We use a simple example to illustrate how this works. Suppose user A (attacker) finds user B (victim)'s whisper in the nearby list, and A wants to pinpoint B's location:

1. A queries the nearby list to get its current distance (d) to victim B (averaged across multiple queries).
2. To estimate the direction, A needs additional observation points. We pick 8 points $\{A_1, A_2, ...A_8\}$ evenly distributed on a circle centered at A with radius d (Figure 24). From each point, A queries the nearby list to measure its distance to victim $\{d_1, d_2, ..., d_8\}$. Suppose X is a dot on the circle, then objective function $Obj = \sqrt{\frac{\sum_{i=1}^{8}(|\overrightarrow{A_iX}|-d_i)^2}{8}}$ reaches the minimum if $\overrightarrow{AX}$ is the right direction to the victim.
3. Then the attacker moves to the next location using $\overrightarrow{AX}$ and d, and repeats step 1 and 2. The algorithm terminates if $d < Thre_1$, or the distance d from two consecutive rounds differs $< Thre_2$.

In practice, the attacker can script all queries with forged GPS values and does not need to physically move.

Distance Error Correction. Finally, we introduce a final step that uses physical measurements to calibrate and add an additional "correction" factor to location data.

We first post a target whisper at a predefined physical location L (on UCSB campus). Then we measure distances to L using the nearby list from a set of observation points, each with known ground-truth distances to L. The ground-truth distance ranges cover from 1 to 25 miles (in 5 mile increments) and again from 0.1 to 0.9 miles (in 0.1-mile increments). At each increment, we use 8 observation points (as specified above) and use each to query the nearby list 100 times. Figure 25 and Figure 26 plot the ground-truth distance versus the measured distance (for 25, 50 and 100 requests per location). For distances greater than 1 mile, we find that our estimates underestimate true physical distance to the victim. Within 1 mile, it clearly overestimates. This mapping between true and measured distance serves as a guide for generating our "correction factor," which is applied to the final estimate.

7.2 Experimental Validation of the Attack

A Single-target Experiment. We first post a whisper at a pre-defined location on UCSB campus as the target (victim). Then we run the attack algorithm starting from distances of 1, 5, 10 and 20 miles away from the victim. Our algorithm takes the average distance over 50 queries per location, and terminates when the estimated distance from consecutive rounds differ < 0.1 mile or when estimated distance < 0.5 mile (based on Figure 26). We repeat each experiment 10 times and test the performance with and without our distance *error correction factor*. Results are shown in Figure 27 and Figure 28.

We make two key observations. First, the algorithm is very accurate. The final error distance, *i.e.* distance from the estimated victim location to the ground-truth location, is only 0.1 to 0.2 miles. With a radius of 0.2 miles, attackers can already effectively identify user's significant points of interest (*e.g.*, home, work, shopping mall) and reconstruct a victim's daily routine using mobility traces [3]. Second, the results show that distance error correction improves algorithm accuracy significantly and reduces the number of iterations needed to determine the victim's location.

Geographically Diverse Targets. To make sure our results are not biased and specific to a single location, we apply the correction factor computed from local measurements (Figure 25 and Figure 26) to carry out attacks in different cities. More specifically, we post target whispers in Santa Barbara and Seattle Washington, Denver Colorado, New York City, New York and Edinburgh Scotland. All whispers are posted via an Android phone with forged GPS coordinates. Then we run the algorithm with distance error correction. We find the final error distances are consistently less than 0.2 miles, and that our correction factor can be generalized to improve estimation accuracy regardless of geographic region.

7.3 Countermeasures

This type of statistical attack cannot be mitigated simply by adding more noise into the system. Attackers can always apply increasingly sophisticated statistical and data mining tools to eliminate noise and determine the true location of a whisper. Instead, the key is to restrict user access to extensive distance measurements. This means putting more constraints (*e.g.*, rate limits) on queries to the nearby list. For instance, one approach is to enforce per-device rate limits. Another is detect fake GPS values, either by relying on client hardware (difficult) or by detecting "unrealistic" movement patterns by potential attackers. Finally, the ultimate defense is to simply remove the "distance" field altogether. While the Whisper engineering team has already addressed this issue, we are not aware of the specific steps they took to do so.

8. RELATED WORK

Online Social Networks. Over the last few years, researchers have performed measurement studies on online social networks (OSNs) including Facebook [36, 39], Twitter [8, 25], Pinterest [12], and Tumblr [9]. Today's OSNs have stored large volumes of sensitive data about users (*e.g.*, personal profile, friending information, activity traces), all of which pose potential privacy risks. Various techniques have been proposed to compromise user anonymity and infer users' sensitive information from social network data [5, 26, 27, 44]. Our study focuses on anonymous social networks, which prioritize user privacy at the cost of eliminating persistent identities as well as social links.

Anonymous Online Communities. Anonymous online services allow users to post content and communicate without revealing their real identity. Researchers have studied various anonymous platforms including anonymous forums [32], discussion boards [6, 23] and Q&A sites [21]. Most earlier works study user communities focusing on content and sentiment analysis. More recently, anonymous social networks have emerged, particularly on mobile platforms. A recent work [31] conducted a user survey on SnapChat to understand how they used the anonymous social app. In comparison, our study is the first to quantitively study user interaction, user engagement, and security implications in the anonymous Whisper network.

Device Localization. Our attack algorithm to localize Whisper users is inspired by existing techniques used for device localization in wireless (mobile) networks [15, 20, 43]. We differ from existing techniques in our approach to deal with the random errors injected by Whisper server. Also, our contribution is more on identifying and validating the security vulnerability instead of the localization algorithm itself.

9. CONCLUSION AND FUTURE WORK

Anonymous, mobile-only messaging apps such as Whisper mark a clear shift away from traditional social networks and towards privacy-conscious communication tools. To the best of our knowledge, our study is the first large data-driven study of social interactions, user engagement, content moderation and privacy risks on the Whisper network. We show that without strong user identities or persistent social links, users interact with random strangers instead of a defined set of friends, leading to weak ties and challenges in long-term user engagement. We show that even in anonymous messaging apps, significant attacks against user privacy are very feasible. We believe that this shift towards privacy in communication tools is here to stay, and insights from our study on Whisper provides value for developers working on next generation systems in this space.

Whisper is not only a social communication tool, but also a network for sharing anonymous content. Analysis and modeling of topics and sentiments in Whisper would be interesting topics for future work. For example, whether and how do users establish communities around "topics" or "themes"? How can anonymous posts and conversations impact user sentiment and emotions? How does user behavior on Whisper compare to those of existing content networks such as Digg and Quora?

Acknowledgments

We would like to thank our shepherd Alan Mislove and the anonymous reviewers for their comments. This project was supported in part by NSF grants IIS-1321083, CNS-1224100, IIS-0916307, by the DARPA GRAPHS program (BAA-12-01), and by the Department of State. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

10. REFERENCES

[1] ALMUHIMEDI, H., WILSON, S., LIU, B., SADEH, N., AND ACQUISTI, A. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proc. of CSCW* (2013).

[2] ANDREESEN, M. Public tweets. Twitter, March 2014.

[3] ASHBROOK, D., AND STARNER, T. Using gps to learn significant locations and predict movement across multiple users. *Personal Ubiquitous Comput. 7*, 5 (2003), 275–286.

[4] ASSOCIATED PRESS. Whispers, secrets and lies? anonymity apps rise. USA Today, March 2014.

[5] BACKSTROM, L., DWORK, C., AND KLEINBERG, J. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proc. of WWW* (2007).

[6] BERNSTEIN, M. S., MONROY-HERNÁNDEZ, A., HARRY, D., ANDRÉ, P., PANOVICH, K., AND VARGAS, G. G. 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proc. of ICWSM* (2011).

[7] BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R., AND LEFEBVRE, E. Fast unfolding of communities in large networks. *JSTAT 2008*, 10 (2008).

[8] CHA, M., HADDADI, H., BENVENUTO, F., AND GUMMADI, K. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. of ICWSM* (2010).

[9] CHANG, Y., TANG, L., INAGAKI, Y., AND LIU, Y. What is tumblr: A statistical overview and comparison. *CoRR abs/1403.5206* (2014).

[10] CLAUSET, A., SHALIZI, C. R., AND NEWMAN, M. E. Power-law distributions in empirical data. *SIAM review 51*, 4 (2009), 661–703.

[11] GARCIA, D., MAVRODIEV, P., AND SCHWEITZER, F. Social resilience in online communities: The autopsy of friendster. In *Proc. of COSN* (2013).

[12] GILBERT, E., BAKHSHI, S., CHANG, S., AND TERVEEN, L. "i need to try this!": A statistical overview of pinterest. In *Proc. of CHI* (2013).

[13] GILBERT, E., AND KARAHALIOS, K. Predicting tie strength with social media. In *Proc. of CHI* (2009).

[14] GONG, N. Z., XU, W., HUANG, L., MITTAL, P., STEFANOV, E., SEKAR, V., AND SONG, D. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proc. of IMC* (2012).

[15] GONZALEZ, M. A., GOMEZ, J., LOPEZ-GUERRERO, M., RANGEL, V., AND OCA, M. M. GUIDE-gradient: A guiding algorithm for mobile nodes in wlan and ad-hoc networks. *Wirel. Pers. Commun. 57*, 4 (2011).

[16] GROVE, J. V. Secrets and lies: Whisper and the return of the anonymous app. CNet News, January 2014.

[17] GUO, L., TAN, E., CHEN, S., ZHANG, X., AND ZHAO, Y. E. Analyzing patterns of user content generation in online social networks. In *Proc. of KDD* (2009).

[18] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *JMLR 3* (2003), 1157–1182.

[19] HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., AND WITTEN, I. H. The weka data mining software: an update. *SIGKDD Explor. Newsl. 11*, 1 (2009).

[20] HAN, D., ANDERSEN, D. G., KAMINSKY, M., PAPAGIANNAKI, K., AND SESHAN, S. Access point localization using local signal strength gradient. In *Proc. of PAM* (2009).

[21] HOSSEINMARDI, H., HAN, R., LV, Q., MISHRA, S., AND GHASEMIANLANGROODI, A. Analyzing negative user behavior in a semi-anonymous social network. *CoRR abs/1404.3839* (2014).

[22] JONES, J. J., SETTLE, J. E., BOND, R. M., FARISS, C. J., MARLOW, C., AND FOWLER, J. H. Inferring tie strength from online directed behavior. *PLoS ONE 8*, 1 (2013), e52168.

[23] KNUTTILA, L. User unknown: 4chan, anonymity and contingency. *First Monday 16*, 10 (2011).

[24] KWAK, H., CHOI, Y., EOM, Y.-H., JEONG, H., AND MOON, S. Mining communities in networks: a solution for consistency and its evaluation. In *Proc. of IMC* (2009).

[25] KWAK, H., LEE, C., PARK, H., AND MOON, S. What is Twitter, a social network or a news media? In *Proc. of WWW* (2010).

[26] MISLOVE, A., VISWANATH, B., GUMMADI, K. P., AND DRUSCHEL, P. You are who you know: inferring user profiles in online social networks. In *Proc. of WSDM* (2010).

[27] NARAYANAN, A., AND SHMATIKOV, V. Robust de-anonymization of large sparse datasets. In *Proc. of IEEE S&P* (2008).

[28] NEWMAN, M. E. Modularity and community structure in networks. *PNAS 103*, 23 (2006), 8577–8582.

[29] NEWMAN, M. E. J. Assortative mixing in networks. *Physical Review Letters 89*, 20 (2002), 208701.

[30] PETROVIC, S., OSBORNE, M., AND LAVRENKO, V. I wish i didn't say that! analyzing and predicting deleted messages in twitter. *CoRR abs/1305.3107* (2013).

[31] ROESNER, F., GILL, B. T., AND KOHNO, T. Sex, lies, or kittens? investigating the use of snapchat's self-destructing messages. In *Proc. of FC* (2014).

[32] SCHOENEBECK, S. Y. The secret life of online moms: Anonymity and disinhibition on youbemom.com. In *Proc. of ICWSM* (2013).

[33] STRAPPARAVA, C., AND VALITUTTI, A. Wordnet affect: an affective extension of wordnet. In *Proc. of LREC* (2004).

[34] STUTZMAN, F., GROSS, R., AND ACQUISTI, A. Silent listeners: The evolution of privacy and disclosure on facebook. *Journal of Privacy and Confidentiality 4*, 2 (2013).

[35] SULER, J., AND PHILLIPS, W. L. The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *Cyberpsy., Behavior, and Soc. Networking 1*, 3 (1998), 275–294.

[36] UGANDER, J., KARRER, B., BACKSTROM, L., AND MARLOW, C. The anatomy of the facebook social graph. *CoRR abs/1111.4503* (2011).

[37] WAKITA, K., AND TSURUMI, T. Finding community structure in mega-scale social networks: [extended abstract]. In *Proc. of WWW* (2007).

[38] WATTS, D. J., AND STROGATZ, S. Collective dynamics of 'small-world' networks. *Nature*, 393 (1998), 440–442.

[39] WILSON, C., BOE, B., SALA, A., PUTTASWAMY, K., AND ZHAO, B. User Interactions in Social Networks and Their Implications. In *Proc. of EuroSys* (2009).

[40] WORTHAM, J. New social app has juicy posts, all anonymous. NY Times, March 2014.

[41] WORTHAM, J. Whatsapp deal bets on a few fewer 'friends'. NY Times, February 2014.

[42] XU, T., CHEN, Y., JIAO, L., ZHAO, B. Y., HUI, P., AND FU, X. Scaling microblogging services with divergent traffic demands. In *Proc. of Middleware* (2011).

[43] ZHANG, Z., ZHOU, X., ZHANG, W., ZHANG, Y., WANG, G., ZHAO, B. Y., AND ZHENG, H. I am the antenna: Accurate outdoor AP location using smartphones. In *Proc. of MobiCom* (2011).

[44] ZHELEVA, E., AND GETOOR, L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proc. of WWW* (2009).