Can you believe an anonymous contributor? On truthfulness in Yahoo! Answers

Dan Pelleg
Yahoo! Research
MATAM
Haifa 31905, Israel
Email: dpelleg@yahoo-inc.com

Elad Yom-Tov †
Microsoft Research
1290 Avenue of the Americas
New York NY 10104
Email: eladyt@microsoft.com

Yoelle Maarek
Yahoo! Research
MATAM
Haifa 31905, Israel
Email: yoelle@ymail.com

***Abstract*—Internet users notoriously take an assumed identity or masquerade as someone else, for reasons such as financial profit or social benefit. But often the converse is also observed, where people choose to reveal true features of their identity, including deeply intimate details. This work attempts to explore several of the conditions that allow this to happen by analyzing the content generated by these users. We examine multiple social media on the Web, specifically focusing on Yahoo! Answers, encompassing more than a billion answers posted since 2006. Our analysis covers discussions of personal topics such as body measurements and income, and of socially sensitive subjects such as sexual behaviors. We offer quantitative proof that people are aware of the fact that they are posting sensitive information, and yet provide accurate information to fulfill specific information needs. Our analysis further reveals that on community question answering sites, when users are truthful, their expectation of an accurate answer is met.**

I. Introduction

Nearly 20 years ago, the cartoonist Peter Steiner coined the adage "On the Internet, nobody knows you're a dog", joking on the fact that anyone could, and still can, make any sort of claim regarding their online identity and are unlikely to be caught lying. But the fact this can be done does not mean everyone is always doing it. Often, people choose to be honest and candid and reveal true features of their identity, providing personal and even sensitive details. This work explores such cases in which users are truthful and the conditions that allow truthfulness to occur. We primarily focus on Community Question Answering (CQA) sites and more specifically, Yahoo! Answers, which, with more than 1 billion posted answers[1] represents a rich and mature repository of user-generated content on the Web.

We consider CQA sites (such as Yahoo! Answers) as a market [1] and derive our definition of truthfulness from mechanism design, in which an auction is said to be truthful (or incentive compatible) if *"each bidder's best strategy is always to reveal her true valuation"* [2]. In the same spirit, we propose here to consider that an asker is truthful if *"she reveals her true needs and intent in the question she asks"*, while an answerer is truthful if *"she answers to the best of her knowledge in the sole goal of satisfying the asker"*. This is different than *reliability* or *quality* of a user, which have been studied before [3], [4]. But to distinguish prior work from ours we make two distinctions. First, our definition is more appropriate for questions that are not necessarily factoids with a single global answer, but may also include interpretation of specific conditions that pertain to a particular asker. Seond, that social-network methods used traditionally are inappropriate for some of our use cases below, which involve ephemeral user identifiers and other identity-obfuscation tactics.

Truthfulness in a market is essential. Qualitatively, if askers assume that answers received to their questions are untrustworthy, they are unlikely to bother posting them. Similarly, answerers will not engage with askers, if they perceive the latter as not alluding to real problems or providing untrue information. Quantitatively, it has been theoretically demonstrated that one should strive to attain equal levels of truthfulness for both buyers and sellers (askers and answerers, respectively) [5]. When this is attained, social welfare, the amount of trade (volume of user engagement) and users' utility functions are maximized. Thus, the ability to measure truthfulness, provide feedback to users on the level of truthfulness, and strive to improve it, are all critical goals for any CQA site.

As noted above, people are known to be untrustworthy in the information they provide in certain situations. A case in point are surveys and peoples' response to them. Literature on response bias in surveys is rich, and documents distorted data on issues such as obesity [6], income, drinking, drug abuse, sexual behavior, and voting [7]. But this is for all modes of conducting a survey, including face-to-face interviews and paper forms. More appropriately for our line of work, some of the factors that were found to promote truthfulness are computerization (as opposed to paper forms) and self-administration (as opposed to interviewing) [8]. In particular, these modes elicit more frequent (and presumably more accurate) self-reports of sensitive behaviors. Moreover, surveys are initiated by an external entity, whereas web activity (at least of the kind discussed here) has an internal motivation. At face value, this would imply that web activity pertaining to survey response

† Work done while at Yahoo!

[1] http://yanswersblog.com/index.php/archives/2010/05/03/1-billion-answers-served/

would be truthful. In practice it is not, and this simple fact calls for a closer examination, which we attempt below.

In certain contexts, users mis-represent themselves in social media. A classic example is dating sites, in which people consistently represent themselves as taller and richer than they really are[2]. In this case, there is a clear incentive, namely to increase one's value in the dating market. Furthermore, misrepresentation is observed also where there is less of a clear benefit to be gained, e.g., in surveys. For example, people in the western world routinely under-report their weight in health surveys [6]. However, when we examined self-reported weight data, gathered from Yahoo! Answers posts, we found a striking match to the national average (see Section IV).

This work attempts to reconcile these conflicting outcomes, as they pertain to the social web. Our hypotheses are that on CQA media:

1) Askers exhibit a high level of truthfulness on personal and sensitive topics.
2) In sensitive scenarios, users take care to hide their true identity by carefully managing their online personae.
3) A sufficient number of answerers behave truthfully enough to meet the needs of the askers.

In order to study truthfulness, we propose to focus first on these areas for which survey-bias studies have verified a greater likelihood that users might hesitate to be truthful, namely personal and sensitive topics such as body measurements, sexual behavior, and income [6], [9]. Our methodology is to mine user-generated texts for patterns to extract facts, which are later aggregated into insights. The information sources we use are Yahoo! Answers, Facebook, Twitter, and Google groups. In particular, we provide quantitative evidence that on CQA sites users are more truthful on these sensitive topics, when compared to their expression on other media. We claim that the reason for this truthfulness lies in the fact that, for the purpose of asking a question, users of CQA sites can hide behind an ephemeral persona and therefore might feel less exposed than in systems such as Facebook, where the visibility of the persona is central. We therefore study here persona management in the context of personal and sensitive topics.

Before continuing to the main body of this work, we make a diversion to note that under our definition, untruthfulness includes the spam and abuse cases, in which users are malicious and driven by social or economic factors rather than genuine needs. Social media, like email, is clearly exposed to spam and abuse. Anti-spam techniques have been widely researched both in the context of email [10] and social media [11], with some work focusing on Yahoo! Answers specifically [12]. Automatic anti-spam tools as well as peer-moderation mechanisms, which allow users to report abuse[3] on both questions and individual answers, do a decent job handling such cases. They could obviously be improved but this is out of the scope of this work.

This paper is structured as follows: In Section II we describe Yahoo! Answers, a popular CQA site, and data collected from it for our work. Section III compares different question types and their appearance in several social media sites. Section IV explores the truthfulness of askers and of answerers. Section V discusses the implications of our findings.

II. Yahoo! Answers Datasets

We consider here askers and answerers of Yahoo! Answers: Askers visit Yahoo! Answers in order to satisfy a variety of information needs, both narrow or complex (*e.g.,"Are there any markets in London that sell old postcards?"*[4]), advice or opinion seeking (*e.g., "How do I persuade my parents to let me use Facebook?"*) or simply a conversation need over a topic they care about (*e.g., "Why do people use sarcasm?"*). On the other hand, answerers are motivated by social reward as well as the playful experience of earning points.

The rules of the site follow. Users post new questions and assign them to a predefined category (*e.g.* "*Diet & Fitness*"). Any signed-in user can post an answer, and earn points for doing so. The asker may designate one of the answers as the "best answer", which increase the number of points awarded to the answerer. If the asker fails to do so within a given time period, the question goes into a voting stage, during which the community of signed-in users can vote for the best answer. Questions that have been assigned a best answer, either by the asker or the community, are considered resolved.

In order to get insights on truthfulness on Yahoo! Answers, we generated a collection of datasets using various extraction and slicing processes. These datasets were generated from data openly available by crawling the Yahoo! Answers site or by calling the public Yahoo! Answers APIs, except for data pertaining to gender, aliases and deleted questions.

Unless specified otherwise, we only examined questions in English, and thus restricted ourselves to a user population dominated by the US and UK. Other qualifiers (such as "lbs" for weight or "$" for income) further focus the data on the US. Our goal was to center on the specific subsets of the site content that would be representative of the various types of information and activities that are central to truthfulness. Sensitive information includes personally sensitive information (e.g., income, anthropometry data [9]), and potentially sensitive behavior (e.g., sexual behaviors [8]), to which we juxtapose personal non-sensitive information and neutral information.

A. Personal information

Body measurements dataset: For anthropometry data, we extracted, via simple term matching, questions in which a user asks for community opinion on his weight, *e.g.,* "*I am a male, 15 years old. I weigh 75kg, and am 180cm tall. Am I fat?*".

[2]http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

[3]An interesting incident occurred in July 2011 on the Yahoo! Answers French site. The report abuse mechanism was itself abused, when a group of users falsely reported abuses and drew the ire of the community. See http://www.yanswersblogfr.com/b3/2011/07/22/internet-nest-pas-un-lieu-de-non-droit/ (in French).

[4]The text of this and other sample questions is taken from actual site content, but details were modified to prevent identification.

Term	Question count	Questions with complete information
Thin	17,541	547
Skinny	6,189	237
Fat	51,988	1,806
Obese	7,353	253

TABLE I
DATASET SIZE FOR ANTHROPOMETRY DATA.

In general, we extracted questions of the form: "Am I *term*", with "term" being one of: "fat", "thin", "skinny", and "obese". We then scanned the text using several regular expressions to extract the age, gender, weight and height of the asker. Depending on the way the question is worded, this process may find only some of the measurements. For example, some askers neglect to specify their gender, or post a photo of themselves instead of giving a textual description. Details on this dataset are given in Table I.

B. Personal sensitive information

Here, we generated two datasets. The first one relates to personal income. The sensitivity of this topic is culturally-specific, but is considered highly sensitive in the USA. The second dataset we considered is one that we termed "secret questions", for which users specifically state they cannot share information with anyone, as detailed below.

Income dataset: We found 592 users who specified their income in a Yahoo! Answers question, and obtained the median income for their ZIP code. A typical example of question in this dataset typically relates to taxes, such as "*I'm a single woman. I make $30.27/hr. I work 12 hour shifts, 5 days a week. What is my actual wage per hour after taxes?*".

"Secret questions" dataset: A more original dataset, which we expected to be very informative on truthfulness issues, pertained to the so-called "secret questions". Following Hasler and Ruthven [13], we identified a set of "secret" questions by matching phrases such as "I cannot tell anyone" and various modifications where the verb in the phrase may match "ask" or "talk to", and the object may match one of a number of close social relations (i.e., "mom"), or professionals (i.e. "my doctor"). Overall, 5,250 questions matched our regular expression patterns[5]. An example of such a secret question is: "*Do I Have A STD? I think I might have an STD I'm 17 and I can't tell my parents.* "

In order to validate this dataset, we modeled the questions' text (minus the regular expression words) using a vector-space model, and compared the distribution of words in the "secret" questions with that of the general Yahoo! Answers corpus. The words that appear with the highest likelihood in the "secret" questions compared to general questions (grouped manually by likely topic) are shown in Table II.

[5]Example patterns include: `I (can't|cannot|could never|will not) (tell|ask|talk to) (anyone|nobody |anybody|mom|mum|dad|my (friends|parents|mom|mum |dad|bf|gf|boyfriend|girlfriend|family|doctor))`.

Topic	Secret Words
Mental	suicidal, depression, therapist, abuse(d), psychiatrist, bipolar, lonely, feel(ing), sad, stupid, mental, memories, guilty, overcome, anger, mood, hate, crying
Family & Friends	uncle, aunt
Sexual	gay, sexually, abortion, bi, yeast, vagina, sex
Others	guidance, myself, cope, yelled, myself, stop, advice, yelling, childhood, anymore, awful, yourself, me, commit, aim.

TABLE II
TOP SECRET WORDS PER CATEGORY.

Category	Examples
Family & Friends	daughter, husband, aunt
First person singular	I, me, mine
Anxiety	worried, fearful, nervous
Sadness	crying, grief, sad
Negative emotion	hurt, ugly, nasty
Sexual	horny, love, incest
Anger	hate, kill, annoyed
Health	clinic, flu, pill

TABLE III
TOP CATEGORIES FOR "SECRET" ISSUES.

In addition, we categorized the words in the "secret" set into about 70 categories using commercial software[6]. We then compared the occurrence counts per category to those in a corpus of general Yahoo! Answers questions. The categories represented more strongly, by a difference of at least 30% in frequency, in the "secret" set as compared to the general set of questions on Yahoo! Answers are shown in Table III.

These findings clearly demonstrate that "secret" questions deal with highly sensitive subjects, including mental, sexual, and social issues.

1) Potentially Sensitive Behavior: In addition, we generated two datasets that could expose potentially sensitive behavior as discussed below.

"Age of first intercourse" dataset: We extracted 66,327 questions which mentioned the words "first time", "sex", and age, in their text. Typically, these questions refer to teenagers who either recently experienced or have concrete plans to experience, first time sexual intercourse with their partners, e.g., "*I am 17 years old. I had sex for the first time ever and it was unprotected. I took plan B. Am I pregnant?*".

"Age of consent" dataset: Age-of-consent questions consisted of 235 questions of the type "*What is the age of consent in New York?*". These questions typically include the ages of two individuals, and discuss the legal aspects of them having sexual intercourse.

2) Neutral dataset: Finally, our last dataset was a smaller one, and relatively neutral in terms of personal and sensitive behavior.

"Quadratic equations" dataset: This set includes 94 questions of the form "*I need to solve this problems[sic]:* $x^2 - 4x - 5 = 0$". They are typical in the "homework help" category, and range in difficulty and in the amount of

[6]`http://www.liwc.net/`

Category	Examples
Ingestion	dish, eat, pizza
Leisure	cook, chat, movie
Time	end, until, season

TABLE IV
TOP CATEGORIES IN FACEBOOK-PRIVATE.

Category	Examples
Swear	(omitted here)
Body	cheek, hands, spit
They	they, their, they'd
Anger	hate, kill, annoyed
Sexual	horny, love, incest
Biological processes	eat, blood, pain
Ingestion	dish, eat, pizza

TABLE V
TOP CATEGORIES IN FACEBOOK-DELETED.

supporting explanations needed for a correct answer. We chose the ones that matched a fairly elaborate regular expression, designed to help us automate the finding of a correct solution.

III. CROSS-MEDIA ANALYSIS

We conducted a cross-media comparison of Yahoo! Answers, Facebook, Google Groups, and Twitter. To build the baseline language models, we first extracted a representative set for each media, as follows. For Yahoo! Answers, we used the API to randomly sample about 140K questions. For Facebook, we used the search API and extracted a sample of 62,740 fresh public wall posts during September 2011 that matched English stop words. For Twitter, we used the approximately 1M messages collected in [14]. From Google groups, we extracted 227 messages that matched stop words.

To build language models for the "secret" corpora, we extracted the following sets. Data from Yahoo! Answers is as described in Section II. Next, we issued the "secret" terms as search phrases to the respective search APIs of Facebook (546 matches), Twitter (117 matches, search restricted to the first three months of 2010) and Google Groups (76 matches). Below, these sets are denoted by the respective medium name, followed by the word "secret".

For Facebook, we created two additional samples: On Facebook, the visibility of wall posts can be set to either public or friends-only. We used the set of posts extracted by an application [15] running under the user's credentials, and consequently able to see friends-only posts from her friends. There were 62,202 such posts. This set is denoted by "Facebook - private".

The second Facebook set was created by attempting to retrieve each of the background messages after a waiting period of about 4 hours for each, and making note of the ones which disappeared from the respective user's wall. There were 3,008 such posts. Posts may disappear in such a way by the user either explicitly deleting them, or limiting their visibility from public to friends-only ("Facebook - deleted" below).

Comparing Facebook-private to the baseline Facebook model, the categories (found as detailed in the previous section) represented more strongly in the private set are shown in Table IV. The categories most highly represented in Facebook-deleted, compared to Facebook, are listed in Table V.

Additionally, we note that searching on Facebook wall posts for income or weight statements using the very same methodology used in Section IV-A yielded virtually no matches.

In summary, we found that information shared with Facebook friends is not particularly sensitive, in the sense that secret Yahoo! Answers deal with much more stigmatic topics, as shown in the previous sections. This is not, however, due to increased ignorance or apathy that would be particular to the Facebook user base. Quite the opposite is true. Facebook users are highly aware of social norms and of their public image, as evidenced by the prominence of swear words, and anger and sexual topics in the posts they later regret and amend, and is further supported by the lack of income or weight information sharing. Therefore, the difference in the types of information shared in both sites is inherent to their respective designs and de-facto codes of conduct.

To create an encompassing visual map of all media under study, we represented each source by a vector-space model of its word probabilities, smoothed using Jelinek-Mercer smoothing with $\lambda = 10^{-2}$. The Jansen-Shannon divergence between each pair of models was computed, and the distances embedded into two dimensions by plotting the first two eigenvectors. The result is shown in Figure 1. The Figure shows that while the non-secret corpora are fairly close to each other, their secret counterparts are further away, and also far away from each other. This means not only that each secret set is separated by language from its public counterpart (which is expected to some extent, given it was chosen by a pattern match), but also that the types of revelations people make in confidence varies by media. An extreme example of this is the separation between "FB private" and "FB secret" (that is, between information marked for friends-only distribution and information marked by a secret phrase, but broadcast publicly). Also note that Google Groups is farther away from the other social media sources, possibly because it is used for discussion-based interaction rather than information sharing more similar to broadcast, as in the other media types. In summary, this supports the conjecture that people carefully choose what, where, and how they convey their personally sensitive information, even when they use notoriously low-privacy channels.

In the interest of completeness, we also align our findings to those of Hasler et al. [13], along two important metrics. The first one is the information need. In that study, this was done by manually reading and classifying about 400 Google groups posts. Since our data is more voluminous, we did this by observing the categories in which the questions appear, and manually classifying them into the same (broad) categories. Results are in Table VI. The second metric is the entity from which the information is hidden. To tabulate this, we used the occurrence counts for each possible object in the pattern

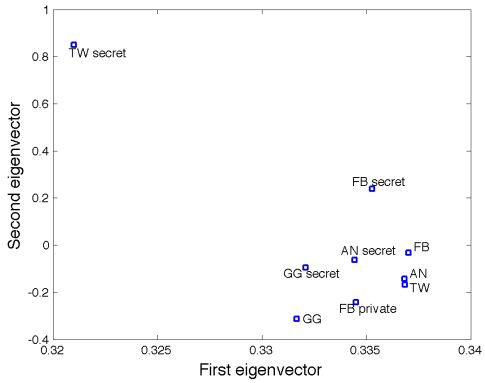

Fig. 1. Distances between the language models of the three social media sources: Yahoo Answers (AN), Facebook (FB), Google Groups (GG), and Twitter (TW).

Yahoo! answers	Need	Google Groups
37.0%	Health-Condition	52.6%
27.5%	Relationships	21.9%
5.3%	Other	
4.0%	Sexuality	3.5%
2.0%	Legal	5.3%

TABLE VI
COMPARISON OF INFORMATION NEED RATES.

we matched ("I cannot tell X"). See Table VII. We observe a good match in the information need, which is noteworthy because of the large difference in the way the two different sites operate.

IV. ASKERS' AND ANSWERERS' TRUTHFULNESS

A. Askers' Truthfulness

We examine here the truthfulness of askers in Yahoo! Answers, restricting ourselves to personal and sensitive questions. Indeed as mentioned earlier, these are domains in which the askers are known to be most reluctant to publicly share information and thus should have the most incentive not to be fully truthful (if we exclude the obvious spam and abuse cases, which are out of the scope of this work). Our goal is to demonstrate that the information provided by askers, when asking questions pertaining to these topics, concurs with corresponding data from other sources. While this method does not formally demonstrate the overall truthfulness of askers on the site, it should still give a good indicator for it.

1) Personal Information: We first focus on personal issues using the body measurement (anthropometry) dataset described in section II. We first discuss three types of personal information that askers typically include in such questions: gender, weight, and height.

Askers often state their gender in questions of this dataset because height and weight norms are typically gender dependent. As we rely on gender to estimate truthfulness on reported weight and height, as detailed below, we compared the gender information given by askers in their profiles, with the one stated in their questions, In our dataset, we had a total of 5,294 users who provided this information in both

Yahoo! answers	Hide target	Google Groups
44.0%	Parents	38.6%
36.7%	Everyone	20.2%
6.9%	Friends	14.9%
5.3%	Family	24.6%
1.6%	Partner	4.2%
1.4%	Professionals	13.2%
0%	Unspecified	9.6%

TABLE VII
COMPARISON OF HIDING TARGET RATES.

questions and profiles. We found that in 96.2% of the cases, there was agreement between the given gender information (statistically significant, χ^2-test, $p < 10^{-10}$). This agreement is approximately uniform (and always above 95%) for all weight-related questions we examined.

Note that these results do not represent a full proof that users are truthful in their profiles. Malicious users can systematically (and consistently) lie and specify a false gender in their questions, as well as in their profile at registration time. Yet, since registering and asking questions are activities that do not typically occur at the same time, this would probably be restricted to a limited number of impersonators. This sanity check was however sufficient for us to use the gender data specified by askers in their questions in conjunction with their self-reported weight and height, when conducting our cross-validation analysis, as detailed below.

We computed the correlation between the average reported weight and height for each age and gender, and the average measured weight and height as given in recent studies, which relied on actual measurements by the Centers for Disease Control and Prevention (CDC) in the US teenager population [16]. The match was extremely high, with an R^2 of 0.97 (0.88) for women (men) for weight ($p < 0.01$) and an R^2 of 0.85 (0.87) for height ($p < 0.01$). See Figure 2. Interestingly, the biggest gap between the curves occurs in women's weight, as their reported weight was, on average, 2.8kg lighter than the US national average (even though the most frequent form of the original question was "Am I fat?").

The match of average height to the values given in CDC data [16] (which is representative of the US population) shows that Yahoo! Answers data is representative of height. The mismatch we observe in weight measurements of women, taken together with the highly correlated trend (across ages) with the trend found in [16], indicats that, while the sample is not representative, it is likely a true indication of asker weight.

Using the data provided by each asker, we computed their Body-Mass Index (BMI)[7] and the percentile thereof, corrected for age according to Ogden et al. [17]. We linked this information to the users' location in the US using their ZIP code. Although this data is known to be noisy [18] it can still be considered a good approximation for users' location.

We modeled the known obesity level for each US county

[7]For adults, BMI is typically calculated as Weight in Kilograms / (Height in Meters)^2. For children, age and gender are taken into account and a BMI percentile is used instead.

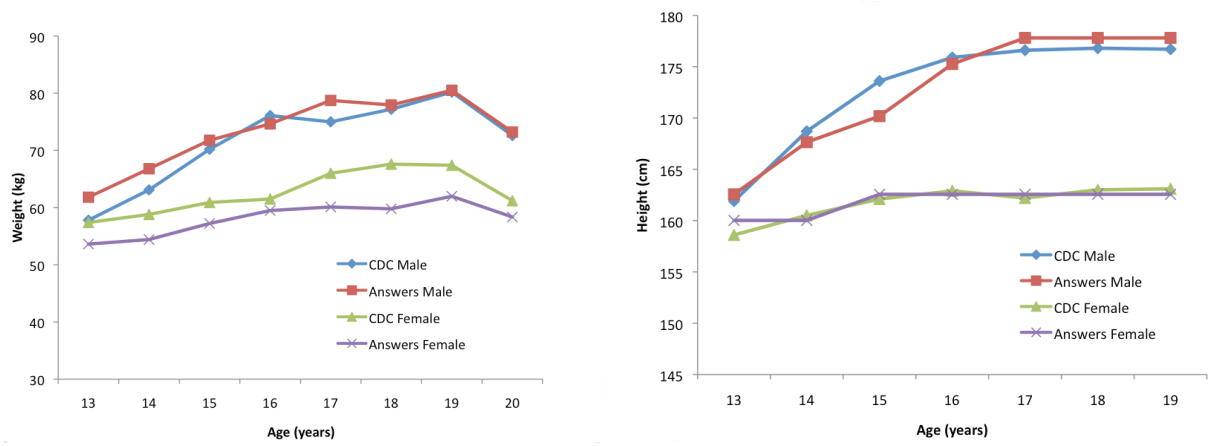

Fig. 2. A comparison of the average weight (left) and height (right) estimated for male and female teenagers from Yahoo! Answers, compared to data collected by the CDC [16].

[19], *i.e.*, the number of people with a BMI greater than 25, using the average BMI per county. Also, we extracted the median household income at that county [20] (as a proxy for Internet accessibility). We found that a linear regression model obtained an R^2 of 0.31, demonstrating that the BMI reported by Yahoo! Answers users concurs with the known geographic variability in the USA.

We discuss below more sensitive issues than the ones above, focusing specifically on income and age of first intercourse, which askers should in general be even more reluctant to publicly share, at least in the western world today.

2) Personal Sensitive Information: We used here the "income dataset" described in Section II. We compared the income figures reported in the dataset questions to the 2001 USA median household income data [20], corrected for inflation. We found that the null hypothesis, that the reported sample and the median income at the users county were from the same distribution, could not be rejected (sign test, $p > 0.05$).

Thus, there is evidence that the individual income reported by Yahoo! Answers users concurs with known population income.

3) Potentially Sensitive Behavior: Finally, we considered questions that pertain to the age of first sexual intercourse, as provided in the dataset described in the previous section. We compared the age indicated in these questions to the age reported in surveys conducted in the USA [21]. We found that the cumulative distributions of ages correlate extremely well ($R^2 = 0.98$, $p < 10^{-4}$, null hypothesis cannot be rejected for the χ^2-test at $p < 0.05$).

The above measurements for both personal/sensitive information as well as potential sensitive behavior show that, in each case, the correlation between real-life measurements of data and their sample extracted from Yahoo! Answers was quite high. This is especially encouraging given that we correlated with survey data that rely on uniform sampling techniques, while there is no guarantee that the askers in Yahoo! Answers form a representative sample of the population. Although it is impossible to safely extrapolate askers' and answerers' truthfulness to the full set of users in Yahoo! Answers, it would appear that, at least for these personal and sensitive topics, users, are truthful in their representation of their characteristics and behavior.

B. Answerer's Truthfulness

We focus here on the truthfulness of the askers' partners in Yahoo! Answers, namely the answerers. As discussed in the previous section, we considered only the providers of "best answers", under the assumption that the original askers and the community already filter irrelevant or blatantly untruthful answerers. The more interesting cases here are those of answerers who might have deceived the asker or the community and got their answers selected as best answers, while being untruthful. We consider here answers that pertain to:

1) a neutral topic, namely solutions to quadratic (second-order) equations, (referred to as topic 1 below),
2) potentially sensitive behavior, namely the age of consent at different locations in the USA (referred to as topic 2 below.)
3) personal information, namely body measurements, (referred to as topic 3 below),
4) personal sensitive information, in its most sensitive form, namely answers to "secret questions", (referred to as topic 4 below).

Evaluating absolute truthfulness of the best answers for questions in these topics is quite challenging. To approximate this, we used correctness, namely that a correct answer would imply a truthful answerer. Note that the converse is not necessarily true, an answerer could be truthful yet incorrect as we will see for topic 3 for instance. We manually labeled all of the questions whenever possible (specifically for all answers provided to questions of type 1 and 2 above and for 1000 of the answers pertaining to type 3). In all these

cases, we found the correct answer and assessed whether the best answer to the question is factually correct. A subset of the questions were labeled by an additional person, and inter-annotator agreement (Kappa statistic [22]) was very high at 0.86. For type 4 above, the so-called "secret questions", we could not use any ground-truth of "correct" answers, as they were mostly opinion or advice seeking questions, so we relied only on askers' feedback as detailed below.

Our results were as follows:

For topic 1, (quadratic equations), which typically require high-school level knowledge of mathematics, 85% of the best answers were factually correct, *i.e.*, gave the correct solutions to the equations. Interestingly, several best answers criticized the asker. For example: "*I think you should figure them out for yourself. When people give you the answer it's called cheating, love*"[8]. We applied a strict criteria and marked those kinds of answers, as well as answers containing just links or explanations on the proper way to solve the question, as incorrect. A more lenient evaluation would have resulted in higher accuracy numbers.

For topic 2, (age of consent), which can be answered by searching the Web (and reading, for instance, the relevant wikipedia page), the percentage of correct answers for age reached 77%. An example of an untruthful answer is: "*You know that there's this thing called google where if you type in Florida age of consent it would tell you*". Here, too, we applied a strict criterion, for example marking answers which failed to mention certain legal nuances as incorrect.

For topic 3, (body measurements, mostly weight-related questions), which typically requires some basic expertise, such as familiarity with the previously mentioned BMI measure, correct answers were given in 74% of the cases if age was not considered, and only in 66% of the cases otherwise. Indeed, askers of weight-related questions are most often teenagers, and if answerers use BMI, the correct usage of BMI requires using age information (See footnote 7). So about 8% of answers were not perfectly accurate but probably not intentionally so, and as per our original definition of truthfulness still truthful. In any case, most of the errors in this topic were due to fuzzy borderlines between various categories. As an example, some of the answerers to questions pointing to a BMI of 25.1 reassured the askers they had a normal BMI, even though, strictly speaking, the accepted cutoff is 25.

These results indicate that, in general, askers receive good factual answers, but their quality degrades as questions become more difficult and require more specialized expertise. Still the majority of answerers remain truthful.

For topic 4, (the "secret questions"), we relied only on the asker's evaluation. We found out that a best answer was selected by the asker in 39% of the cases, compared to 32% in a control group[9] (statistically significant, χ^2-test, $p < 10^{-6}$). This indicates that askers were more content with the answers compared to the general population. Yet results are clearly less conclusive here as our approximation of accuracy for truthfulness can hardly be used, the topic being more subjective and open-ended. Additionally, answerers of "secret" questions, in comparison to the general user base, are active in fewer categories (27.7 vs. 30.4) and their answers were chosen less frequently as best answers (603 vs. 830) (t-test, $p < 0.01$ in both cases). This might imply that they do contribute to Yahoo! Answers in order to maximize their score, but rather have a genuine interest in a particular ("secret") subject.

C. Persona management

Given the sensitivity of some of the questions we considered in our datasets, we wanted to investigate whether askers are comfortable exposing their public persona, while discussing these topics. We used for this purpose the previously mentioned so-called "secret questions dataset" as it clearly exposes public visibility concerns, on the asker's side.

In Yahoo! Answers, users can control the visibility of their persona by choosing to hide aspects of their activity from public view. We found that 23% of the askers of "secret" questions hide their list of friends from public view, compared to just 17% in a sample of all users, and that 34% of "secret" askers hide both their questions and answers from public view, compared to just 23% in the control group (χ^2-test, $p < 10^{-10}$ in both cases). While these values may seem low, consider that the default option is full visibility for both of them. The power of the default option has been well documented, even in the context of privacy [23], therefore any change against the default is significant.

Users can also describe themselves using two textual fields: a nickname and a resume. We found that only 17.3% of "secret" askers provide a resume, compared to 18.2% of the control group. Similarly, 4.3% of "secret" askers choose a nickname that is identical to their Yahoo! identifier, compared to 5.3% of the control group[10]. Thus, "secret" askers are more likely to mask their identifies.

Moreover, askers can actively choose to "cover their tracks" in two different manners. They can either create a unique user for sensitive questions or delete the questions once they receive an answer to them. We have found that askers of "secret" questions use both strategies, as we discuss below.

1) Using multiple personae: When examining the user names who asked "secret" questions, we found that they ask, on average, only 1.2 questions, which is significantly lower than the average 5.3 questions per user in a control group. They also very rarely answer questions, namely 1.1 on average as compared to 28.6 for the control group ($p < 10^{-10}$ in both cases). Our interpretation here is that frequent users of the site, who already have a public persona there, will define a new alias to ask a "secret" question. Alternatively, if they

[8]Remember that in some cases the best answer is chosen by the community rather than by the asker.

[9]We chose a control group by drawing a sample of 5,172 questions, stratified to match the category distribution of the "secret" questions.

[10]Anecdotally, the vast majority of nicknames are not reflective of actual person names, and the images which could be used to show them are overwhelmingly either the default icons or images which are unrelated to the user (e.g., celebrities).

are casual users, they will come to Yahoo! Answers to ask one very sensitive question and will not come back for other activities.

Furthermore, there are indeed quite a few users who maintain more than one alias on Yahoo! Answers, though they have a single Yahoo! user name. 3,709 such users were identified, after limiting our analysis to users who posted at overlapping times using at least two user names (to exclude people who forgot their original user name and registered a new one instead). We define here as the *majority persona* the user name through which most questions were made and as the *minority persona*, the next one in frequency. Note that 97% of users with more than one user name had exactly two user names, therefore we did not consider less frequently active user names.

We computed the probability of posting in each Yahoo! Answers pre-defined category using the minority persona and using the majority persona, and computed the categories for which the ratio between the two is highest. We list below the ten categories for which this ratio is greatest, grouping them by the type of information or behavior as defined in defined in Section II. The parenthesized numbers are the ranks, 1 indicating the highest ratio between probabilities, and 10 the lowest.

1) Personal information: "Hair" (3), "Diet and Fitness" (6).
2) Personal sensitive information: "Lesbian, Gay, Bisexual, and Transgendered" (1), "Adolescent" (5), "Mental Health" (7).
3) Potentially sensitive activity: Yahoo! Answers (2) (gaming the Answers system), "Computers" (9) (illegal downloading), "Video and Online Games" (10) (illegal downloading).
4) Others: "Comics and Animation" (8), "Cell Phones and Plans" (4).

Thus, it appears that users who post questions using their minority persona do so to hide questions that are either personal or sensitive, while using the majority persona for other topics.

2) Deleting questions: Another way to manage one's public persona is simply to delete the question once the asker realizes she does not want this question to be associated with her persona, this option is available to all askers. Clearly user-initiated deletion can be caused by other reasons, such as having the information need satisfied from other sources. Yet one can assume that some of these deletions are motivated by regret.

The list below shows the categories with the highest question deletion rate, grouped again by type. The parenthesized numbers are the ranks, 1 representing the highest deletion rate, and 15 the lowest.

1) Personal information: "Polls and Surveys" (2), "Religion and Spirituality" (3), "Beauty and Style" (4), "Fashion and Accessories" (14).
2) Personal sensitive information: "Singles and Dating" (1), "Friends" (6), , "Family and Relationships" (9), "Mental Health" (10), "Lesbian, Gay, Bisexual, and Transgendered" (11), "Psychology" (12), "Women's Health" (13), "Adolescent (parenting)" (15).
3) Potentially sensitive activity: "Mathematics" (5), "Homework Help" (7), "Video and Online Games" (8).

Thus, we have shown that on Yahoo! Answers, users are consciously using methods for managing their online persona to allow them to post sensitive questions while being truthful. This stands in contrast to the way such topics are generally "taboo" on other social media, as discussed in Section III.

V. Discussion

Our analysis in this paper has demonstrated that, in the right setting, Web users exhibit a high level of truthfulness, even when dealing with personal and sensitive topics. This stands in contrast to the well-documented response bias effect in areas such as anthropometry. Indeed, we extracted 7,218 responses from a US government public-health survey [24] that contained information on self-reported age, gender, and weight, and verified that for most age groups in which this is statistically significant, the self-reported weight is lower than the known US average, except for the $80+$ age group, where it is higher[11]. We found that gender and survey formats or medium are also significant: males only mis-represent themselves when interviewed by phone, while females do it by either phone or mail.

So it appears that the very act of volunteering the data for online sharing, also improves its accuracy — unless the author has something to gain. This ties to a result by Raban [1], demonstrating that Yahoo! Answers is an Online Information Market, where users are motivated by a mix of social and economic incentives. It follows that understanding whether and when users are truthful is critical not only to the proper functioning of the market but also for derivative applications such as displaying the right ads on the site, or spotting fraudulent auction items.

We hypothesize that there are several reasons for which the level of truthfulness is especially high on CQA sites. First, as opposed to surveys, activity on CQA sites is internally motivated, and does not stem from an external entity probing certain subjects. The second reason is anonymity, which has been reported as having a bias-reducing effect [8]. In the context of the Web, this means a disconnect from the real-life persona, and ties directly to our findings on the different types of personae that people exhibit on Facebook and Twitter, as compared to Yahoo! Answers.

The statement above should not be construed as a statement on the different populations using these media; it may very well be the case that some Dr. Jekyll, who freely shares some aspects of his social life on Facebook, occasionally turns to Yahoo! Answers to post a personal question as Mr. Hyde.

We also found that the expectation of a truthful answer is met, at least to the extent of our (admittedly limited) ability to verify the answers. An open question is whether

[11]For this age group, the finding is statistically significant only for males.

one can automatically identify untruthful posts by askers and answerers. We started exploring ways to do just that, and are making progress on this front but the results are not yet ready for publication.

We note the large investment currently needed to conduct large-scale surveys on matters of sensitive medical or personal details. Compare that with our demonstration of obtaining similar data online at much lower cost. The implication is that data collected in such ways has great promise for public policy research. But this is a prospect for the future, after important questions regarding data quality are resolved.

One possible explanation to the differences in data quality, which we alluded to, is the lack of benefit to be gained by misreporting. However, this is not a sufficient reason. For example, we observed gender and other differences in the secretive questions[12], which imply there are deeper psychological processes of self-selection at play here. Such discrepancies limit the level of confidence one can place in far-reaching conclusions drawn from data collected using our method. However, we maintain that the analysis and insights offered here represent a first step toward a deeper understanding of users and their behaviors in both the physical and online worlds, opening directions for a possibly new field of quantitative web sociology.

VI. Conclusions

In this work, we attempted to quantitatively explore truthfulness in social media, with a focus on Yahoo! Answers. To the best of our knowledge, this is the first time such an attempt is made.

Our findings indicate that askers generally provide accurate information, even for highly sensitive topics. Furthermore, askers seem to be aware of the sensitive nature of their questions, and thus they typically go to some length to try and hide their identity when they ask about sensitive topics. When hiding ones' identity is difficult or impossible, such as on deeper reputation systems, like Facebook, the sensitive topics are taboo. Finally, we show that answerers also exhibit high degree of truthfulness, though it varies with the difficulty of the actual subject matter.

We believe this work opens a door to a new line of research, which explores the mechanisms that govern sharing of non-public information on social media. This will lead to a better understanding of not just the way people choose to share pieces of information, but also of the reasons and roadblocks to hide information, with close ties to the psychology of information hiding. Given that truthfulness is a crucial part of CQA systems, such research is needed if CQA sites are to continue serving the needs of users. Future research will demonstrate how providing feedback on truthfulness, improving truthfulness, and automatically detecting it, modify user behavior on CQA sites.

In addition, this work supports research in public health using cheap and publicly-available data, an area which has recently been gaining popularity. While prior work focused on disease outbreaks and rudimentary risk factor analysis [25], we observe that analysis can be done on a much finer granularity and in a much broader set of topics, such as anthropometrics, economics, post-diagnosis patient behavior, and more, some of which we are in the process of authoring articles on.

[12]For example, consider the lack of appearance of "girlfriend" as the person the asker hides from in our "secret dataset", as opposed to the numerous occurrences of "boyfriend".

Acknowledgments

We are deeply grateful to Steven Soria Jr. for kindly making the Facebook data he gathered, while at Stanford, available to us and to Chris Manning for facilitating this. We thank Idan Szpektor for helping with text and content analysis in general and secret questions validation in particular, and Ingmar Weber for pointing us to demographic data. Finally, we thank Prabhakar Raghavan for coining the term quantitative web sociology and attracting our attention to it.

References

[1] D. Raban, "Self-presentation and the value of information in Q&A web sites," *JASIST*, vol. 60, no. 12, pp. 2465–2473, 2009.

[2] A. Archer, C. Papadimitriou, K. Talwar, and E. Tardos, "An approximate truthful mechanism for combinatorial auctions with single parameter agents," *Internet Mathematics*, vol. 1, no. 2, pp. 129–150, 2003.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high quality content in social media, with an application to community-based question answering," in *Proceedings of ACM WSDM*. Stanford, CA, USA: ACM Press, Feb. 2008, pp. 183–194. [Online]. Available: http://videolectures.net/wsdm08_castillo_fhqc/

[4] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha, "Learning to recognize reliable users and content in social media with coupled mutual reinforcement," in *Proceedings of the 18th international conference on World wide web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 51–60. [Online]. Available: http://dx.doi.org/10.1145/1526709.1526717

[5] S. Brainov and T. Sandholm, "Contracting with uncertain level of trust," in *Proceedings of the 1st ACM conference on Electronic commerce*, ser. EC '99. New York, NY, USA: ACM, 1999, pp. 15–21. [Online]. Available: http://doi.acm.org/10.1145/336992.336998

[6] V. Dauphinot, H. Wolff, F. Naudin, R. Guéguen, C. Sermet, J. M. Gaspoz, and M. P. Kossovsky, "New obesity body mass index threshold for self-reported data," *Journal of Epidemiology and Community Health*, vol. 63, no. 2, pp. 128–132, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1136/jech.2008.077800

[7] W. M. Epstein, "Response bias in opinion polls and american social welfare," *The Social Science Journal*, vol. 43, no. 1, pp. 99–110, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0362331905001023

[8] K. E. Schroder, M. P. Carey, and P. A. Vanable, "Methodological challenges in research on sexual risk behavior: II. accuracy of self-reports." *Annals of behavioral medicine*, vol. 26, no. 2, pp. 104–123, Oct. 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/14534028

[9] M. K. Vernon, "Pre-testing sensitive questions: Perceived sensitivity, comprehension, and order effects of questions about income and weight," American time use survey papers and publications, 2005. [Online]. Available: http://stat.bls.gov/osmr/pdf/st050090.pdf

[10] G. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, no. 4, pp. 35–455, 2008.

[11] C. Castillo and B. D. Davison, "Adversarial web search," *Foundations and Trends in Information Retrieval*, vol. 4, no. 5, pp. 377–486, 2010.

[12] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "A few bad votes too many? towards robust ranking in social media," in *Proceedings of the Fourth International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*. New York, NY, USA: ACM, 2008, pp. 53–60.

[13] L. Hasler and I. Ruthven, "Escaping information poverty through internet newsgroups," in *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM-11)*, July 2011, pp. 153–160.

[14] "WWW 2009 workshop on content analysis in web 2.0 (CAW 2.0)," 2009. [Online]. Available: http://caw2.barcelonamedia.org/

[15] J. K. Ahkter and S. Soria, "Sentiment analysis: Facebook status messages," stanford CS224N final project. [Online]. Available: nlp.stanford.edu/courses/cs224n/2010/reports/ssoriajr-kanej.pdf

[16] M. A. McDowell, C. D. Fryar, C. L. Ogden, and K. M. Flegal, "Anthropometric reference data for children and adults: United states, 2003–2006," *National Health Statistics Reports*, vol. 10, 2008.

[17] C. L. Ogden, R. J. Kuczmarski, K. M. Flegal, Z. Mei, S. Guo, R. Wei, and C. L. Johnson, "Centers for disease control and prevention: 2000 growth charts for the united states: Improvements to the 1977 national center for health statistics version," *Pediatrics*, vol. 109, pp. 45–60, 2002.

[18] E. Yom-Tov and F. Diaz, "Out of sight, not out of mind: On the effect of social and physical detachment on information need," in *34th ACM Conference on Research and Development in Information Retrieval (SIGIR 2011)*, Beijing, China, 2011.

[19] "County-level estimates of obesity—US maps," 2008. [Online]. Available: http://apps.nccd.cdc.gov/DDT_STRS2/NationalDiabetesPrevalenceEstimates.aspx?mode=OBS

[20] "U.S. census of population and housing 2000," 2001. [Online]. Available: http://factfinder.census.gov/

[21] W. D. Mosher, A. Chandra, and J. Jones, "Sexual behavior and selected health measures: Men and women 15–44 years of age, United States, 2002," 2005. [Online]. Available: www.kinseyinstitute.org/resources/FAQ.html

[22] J. Cohen, "A coefficient of agreement for nominal scales." *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[23] R. Gross and A. Acquisti, "Information revelation and privacy in online social networks," in *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, ser. WPES '05. New York: ACM, 2005, pp. 71–80. [Online]. Available: http://doi.acm.org/10.1145/1102199.1102214

[24] D. Nelson, G. Kreps, B. Hesse, R. Croyle, G. Willis, N. Arora, B. Rimer, K. Vish Viswanath, N. Weinstein, and S. Alden, "The health information national trends survey (HINTS): Development, design, and dissemination," *Journal of Health Communication*, vol. 9, no. 5, pp. 443–460, 2004. [Online]. Available: http://hints.cancer.gov

[25] M. J. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *ICWSM*, L. A. Adamic, R. A. Baeza-Yates, and S. Counts, Eds. The AAAI Press, 2011.