Touch Screens for Touchy Issues: Analysis of Accessing Sensitive Information from Mobile Devices

Dan Pelleg
Yahoo! Research, Haifa, Israel
pellegd@acm.org

Denis Savenkov
Emory University, Atlanta, GA, USA
dsavenk@emory.edu

Eugene Agichtein
Emory University, Atlanta, GA, USA
eugene@mathcs.emory.edu

Abstract

Smartphones, and other similar devices, are ideal for private activities: They are carried on the body and can be physically secured from prying eyes. They work anywhere. And they are not typically shared among family members or classmates, making them perfect for searching and exploring sensitive information — both via web search or through social information seeking. This paper validates these intuitions empirically, and builds on these observations to investigate the expression of sensitive information needs, such as inquiring about unplanned pregnancies, among users of mobile devices. We examine a large set of queries from the United States (4.2 million), submitted to Yahoo! Answers, and compare these to the 6.7 million queries submitted to Yahoo! search. Focusing on community question answering allows much richer analysis of the way information needs are expressed, setting our work apart form previous studies. For the first time, we empirically show that people prefer to express sensitive needs on mobile devices, as this manifests at both the lexical level and the semantic level, using a pre-defined taxonomy of topics. Further, we find that preference for mobile devices for sensitive topics holds true even when controlling for age and gender, which is facilitated by the large sample of users in our study (1.5 million). In particular, we show that young users, especially females, are more likely to inquire on sensitive issues, and are more likely to do this from locations distant from their common place of web access. To our knowledge, this is the largest study to date of mobile social information seeking, and is unique in terms of demographic diversity, the granularity of information needs examined, and the analysis of the location-dependence of sensitive information needs.

Introduction

Smartphones, tablets, and other mobile devices, had become ubiquitous, and are overtaking desktop PCs in popularity, especially with younger users. Additionally, these hand-held devices are ideal for private activities: they are carried on the body and can be physically secured from prying eyes. They work anywhere. And they are not typically shared among family members or classmates, making them perfect for seeking answers to sensitive information needs, both through automated web search engines and through forums and other social sites.

For example, consider the following scenario, which unfolds hundreds of times a day in United States and around the world. A female teenager is using the computer in her school library, or perhaps in her family's living room. She needs to inquire about something sensitive — say, her health, or her high-school romance. She then realizes that classmates or family members might peek over her shoulder while she posts the query, or perhaps later look at the browsing history on the computer. Will she decide, instead of using a desktop PC, to post the question from the smartphone in her pocket? If she does so, will she first retreat to a more secluded space? And, will she choose to use an automated web search engine or a social service, like Yahoo! Answers, to obtain information or seek support? These questions motivate the work in this paper.

To understand the difference between mobile and desktop information needs, we use the data from Yahoo! Answers, a popular community question-answering (CQA) site, as well as from Yahoo! search logs. We collect millions of queries from each source. Since mobile web use is heavily correlated with age (and, as we show below, gender), we use the size of the corpus to carefully control for these effects. We investigate the correlation of mobile usage vs. sensitive topics such as health, sexual orientation, and relationships. We do this at multiple levels: (1) The lexico-semantic level, by comparing terms and topics, (2) The semantic level, by category hierarchy of the CQA site, and (3) At the level of a specific sensitive need, for example, regarding a concern about a potential pregnancy. Our data shows that for all levels of granularity, there is a strong preference for using a mobile device for such sensitive topics.

To better understand the interplay between information need and physical location, we use geo-location by IP address, which generates a mapping from a query into coarse coordinates. We analyze it at the micro level, to generate a per-user list of frequent and infrequent locations. We show that when posting a sensitive question, the asker is more likely to be physically distant from her usual place of web access.

Our specific research questions are:

RQ1 When inquiring about sensitive issues, are mobile devices preferred over desktop computers:

(a) while posting a question on a CQA site?

(b) while issuing web search queries?

RQ2 Is there a relationship between the physical location from where a question or query are submitted, and the sensitivity of the query?

The study reported in this paper is both unique and timely. While previous studies of mobile search information needs were informative (Yi, Maghoul, and Pedersen 2008; Church and Smyth 2009; Teevan et al. 2011; Ghose, Goldfarb, and Han 2010), these studies were performed at much smaller scale than ours, or were done in the early days of mobile computing, before mobile devices gained present-day capabilities, and have become ubiquitous. Smartphones and tablets are not used anymore by only a small group of early-adopters, but are now widely spread over a sizable fraction of the U.S. population. This includes youngsters using web-capable music players to work on their homework at the local coffee shops and cafes (Troianovski 2013), grandparents using tablets at home, and everyone in between.

Therefore, this work fills in an important knowledge gap about mobile web usage, specifically focusing on *information seeking*. The contributions of this paper are:

- A large-scale quantitative analysis of mobile information needs, as expressed in both CQA postings and web search queries.
- Analysis of the relationship between sensitive information needs and mobile devices.
- Detailed demographic analysis of mobile information needs, including a focus on the under-20 age group.
- A methodology to identify how different locations relate to different types of information needs.

In addition to providing a better understanding of the social information seeking behavior in the mobile vs. desktop contexts, our findings could potentially benefit not only social question answering, but also other online services.

Background

In the early days of mobile connected computing — and even as recently as five years ago — mobile web search was considered a novelty and was accessible only to early adapters of the technology. Yi, Maghoul, and Pedersen (2008) find an abundance of adult search terms, and likens the situation to the time when internet search just took off, circa 2000. Then, too, adult-related searches were dominant, and later waned as more diverse populations started using search. In relation to our work, they hypothesize that a mobile device is susceptible to porn searches because the lack of caching (due to insufficient memory) provides better guarantees of privacy.

Mobile searches were found to be associated with time and place. In a 20-user diary study, Church and Smyth (2009) report that 67% of mobile search queries are posted while in transit, that 30% of the needs expressed are related to geography, and that 8.4% of the queries include an explicit temporal need. Teevan et al. (2011), in a survey of 929 IT workers using their work smartphone, matches the rate of in-transit queries (68%). Ghose, Goldfarb, and Han (2010), in a study of 260 users, suggest that the portability of the device gives easy access to timely information, and show that the cost of acquiring timely information is lower on a mobile device, when compared to a PC. In a large quantitative study, Wang, Huang, and White (2013) show that for search tasks that span a device boundary (and in particular start at the desktop and end while mobile), the top pre-switch search needs are navigational (followed by image and then local); the post-switch continuation needs are also topped by navigational needs, and followed by image and celebrities.

Mobile activity is simultaneously (and perhaps contradictorily) both social and private. Teevan et al. (2011) report that 63% of the searches are collaborative, or multi-party, searches. These typically involve planning meetings such as group dinners, with some of the other parties present at the time of search — at times helping with the search task by either using their own phone, or even taking control of the survey respondent's phone.

Ahern et al. (2007) explores the issue of location-based privacy in the context of photo sharing, in a study of 350 users. They report that location while posting is indicative of life patterns. They also make the distinction between frequent and infrequent locations, citing a user claiming that "some locations are more private than others". They find weak support to the hypothesis that photos are more likely to be publicly shared when the location is frequent (like at a user's home). This seems to contrast with our findings, however the needs are different — in CQA, one does not normally broadcast the post to his social network, nor is the poster's location or even modality available to the readers of the post. In a small qualitative study by Mancini et al. (2009), the notion of a *space* is discussed, as a characteristic for a private surrounding in terms of mobile activity. Even for IT professionals who use the phone almost exclusively for work, the smartphone is reported to be "discreet" (in meeting settings) (Karlson et al. 2009).

The work by Lee et al. (2012) is closest to ours, in that it discusses mobile Q&A, and analyzes a large set of 2.4M questions on Naver mobile. They follow up their quantitative analysis with a 555-participant survey. They find that on mobile CQA, users tend to seek more factual information. When they do ask personal questions, the answers are more likely to be opinions than hard information. The top reasons listed by users for choosing mobile CQA over web search were (1) to save time and (2) to receive personalized answers. Of the survey respondents, 5% reported that they preferred CQA over asking their friends because the questions were private or embarrassing. We assume that this number is under-reported, as users had the option to ignore the survey (as did 95% of the users approached), and users embarrassed about their postings would be more likely to do so.

Demographic and geographic information on mobile users in the works above is thin. Many of the papers are small diary studies or surveys, where the users tend to be IT professionals — that is, well-educated, over 25 years of age,

and often workaholics. This introduces significant bias (in our data, 70% of the posts are from people under 25). Other works (Lee et al. 2012; Yi, Maghoul, and Pedersen 2008; Wang, Huang, and White 2013) cover a large population, but do not report any demographic or geographic analysis below the country level. In this sense our work is unique.

In a different line of work, geographic attributes were exploited to expose trends and events in disease outbreaks (Paul and Dredze 2011), weather (Kıcıman 2012) and even earthquakes (Sakaki, Okazaki, and Matsuo 2010). The focus there is on user's reaction to exogenous events, or rather to the aggregated reaction of the crowd, while we look at the intrinsic needs of individuals.

Sensitive information is also hard to come by. With the exception of Yi, Maghoul, and Pedersen (2008) pointing out adult searches, none of the other studies touches the issue. This is not surprising for the diary studies and other work based on self-reports. Pelleg, Yom-Tov, and Maarek (2012) and Hasler and Ruthven (2011) explored the issue of sensitive questions, and found that users will post those, when given control over the level of personal exposure. In particular, users of Yahoo! Answers will post many kinds of sensitive questions, but typically from a throw-away account or after unlinking any personally identifiable information. Thus, our analysis of such sensitive information needs is unique in that (while the user's identity remains anonymous), we are still able, for the first time, to analyze the users' information seeking behavior along demographic, topical, and geographic dimensions at large scale.

In the late 1990s, as mobile phones just started gaining popularity, researchers turned their attention to the new phenomenon. Palen, Salzman, and Youngs (2000) studied the behavior of 19 new mobile phone users. The study found that people initially acquired cell phones for safety, security and "business" reasons. However, in practice they were typically used for social interaction. Wei and Leung (1999) explored the issues of social use of mobile phones in public spaces. They found that personal uses, e.g. calling a family member or a friend, was much more common than non-social uses, e.g. business calls. Aoki and Downes (2003) focused their study on young people and their practices of using mobile phones. Several participants of the study described the use of cell phones to maintain or manage privacy. They used the landline numbers for certain business transactions and kept the cell phone numbers to those who are their in-group members. Below, we show that these behaviors are also reflected in web usage from a mobile phone. Note that we observed significant amounts of content coming from devices without cellular audio capability (e.g., models of Apple iPods, which are music players, but with touch screen and Wi-Fi connectivity). Therefore, the usage patterns are likely to be a function of the form factor and personal nature of the device, rather than a result of the ability to place and receive calls.

Analysis

Data Sets

For our analysis, we used data from two sources: web search and question posting. Web search is a more casual type of activity, requiring little commitment from the searcher, and typically built into the device's default interface. On the other hand, posting questions to a CQA website like Yahoo! Answers, requires locating a particular web site, formulating a question, choosing a category for it, and "paying" for it in the form of giving away game points[1]. In Yahoo! Answers the category is chosen from a pre-defined taxonomy of about 1700 categories (maximum depth 4), with the help of a suggestion from an automatic categorizer. From the search logs, we recorded the query string, and from the question posts we had access to the question subject and body, as well as the category.

Even though the web search activity and the question posting activity were observed during the same period, the set of users is separate (up to random matches). In particular, the search set was heavily down-sampled from the original logs. An alternative approach would be a within-subject design, and while this is technically possible, it would produce much sparser data. We decided to leave this kind of study for future work: all the results below report on the (largely) separate two sets of data.

We used the data from logged-in users only, after anonymization. This includes each user's age, gender, and zip code, as volunteered at the time of registration. For each access we recorded the timestamp and the type of device, as determined by the user-agent string. This can typically identify a particular device (make and model) for all major types of smartphones, tablets, and web-capable PDAs and music players. It also identifies desktop computers, but all makes and models are lumped into a single type, which also includes laptop computers. We also used the IP address of the request, geo-located using a commercial product, to the level of a zip code.

Data was collected between July 12th and August 20th, 2012. We sampled $6.7M$ web searches from Yahoo! search ($2.6M$ after removing duplicate queries by the same user), and $4.2M$ questions from Yahoo! Answers ($3.8M$ of which included zip codes identified by geo-locating the IP address). The basic statistics are shown in Table 1. We see that in terms of character counts, mobile queries are shorter than their desktop counterparts, for both search and question subjects. Interestingly, the lengths reported here are higher than the mobile search numbers 2.35 words and 13.73 characters per query observed in 2008 (Yi, Maghoul, and Pedersen 2008). We also notice that in terms of words per question body, mobile questions are in fact longer, even though they have fewer characters (so the average word length is higher). The majority gender is male for desktop web searches but female for mobile web searches. In question posting no such change is observed — males are the majority in both.

While some observations are expected (like shorter posts when entered from mobile), the actual correlation is not eas-

[1]The site awards points to answerers, but a new user may ask up to 20 questions before his or her initial credit runs out.

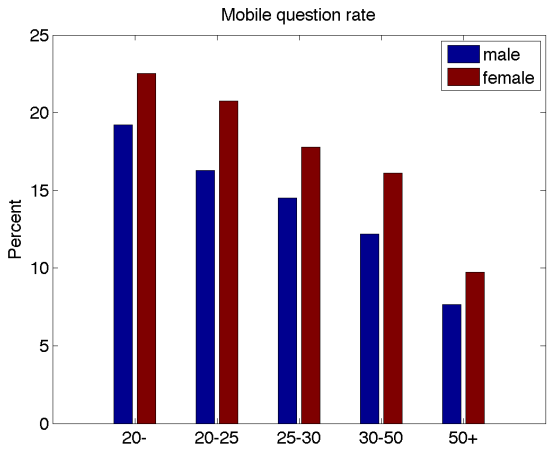


Figure 1: Rates of posting from a mobile device, by age group and gender

ily understood. For example, as younger people are generally more active on the CQA site using a mobile device, it is not clear if youngsters' typing habits are a confounding variable here. In other words, if young people generally use shorter words, and are also a sizable demographic in our data, then short input strings from mobile devices could simply mean that younger people are moving the average, and not that the interaction with the device in general is changing typing habits. To discount these effects, in the analysis below we partition our data by age groups. Where appropriate, we further partition by gender.

Figure 1 shows that age and gender are indeed correlated with mobile usage. While 21.8% of the questions come from mobile devices, this number varies between 7.6% for males over 50 and 22.5% for females under 20. It is consistently higher for females than for males when conditioning by age group, and drops rapidly with age. Furthermore, our age group distribution is far from uniform, with 41.8% of questions coming from the under-20 group, 28.4% from 20–25, 9.1% from 25–30, 15.0% from 30–50, and 5.6% from above 50. Compare this with the mobile-only site reported by (Lee et al. 2012), with 74% teenagers and 15% in their twenties.

To analyze the topics of questions posted from different devices, we computed the occurrence frequencies of each top-level category, when conditioning on each type of device (mobile and desktop). For each category this produces two numbers: its occurrence frequency considering just mobile posts, and its occurrence frequency considering just desktop posts. There are 26 top-level categories for questions and Figure 2a shows the difference in the conditional frequencies, where positive values mean that the mobile rate is higher than the desktop rate, and vice versa. Results show that health, family and pregnancy related questions have higher frequency on mobile devices, but politics, computers, science and culture questions are asked more often from desktop devices. We keep in mind the age skew described above, which could act as a confounding variable. To eliminate a potential bias we look at category distribution for users of particular ages and gender. We calculated conditional frequencies as before, but only for the 20–24 age group, separately for males and females. The results are presented on Figures 2b and 2c. As one can see, mobile and desktop category frequencies for differ by gender, but in both cases questions posted from mobile devices tend to be about personal and sensitive topics. For example, health is still the top "mobile" category, but "Pregnancy & Parenting" moves from third place in the overall population, to second place for females, and to sixth place for males. These results give support for a positive answer to our research question RQ1a.

The analysis of other age groups (data not shown) shows similar results, but also reflects the interests shift by age. For example, for users 30–50 years old, politics and society topics become more popular (both for male and female users) and these categories are asked with a higher rate from desktop devices. But health, relationships, parenting and other personal questions still have higher frequency of being asked from mobile devices.

The 26 top-level categories are very broad in terms of the topics they cover. For example, the "Society & Culture" category has both a "Royalty" sub-category (not sensitive), but also a sub-category for "Lesbian, Gay, Bisexual, and Transgendered" (sensitive). To disambiguate, we drilled down into the leaf categories. For each leaf category and demographic, we counted the number of questions from either type of device, and computed the ratio of questions posted from mobile devices within that category. We then sorted to get the categories that host the highest ratio of mobile questions. See Table 2. Several potentially sensitive categories emerge:

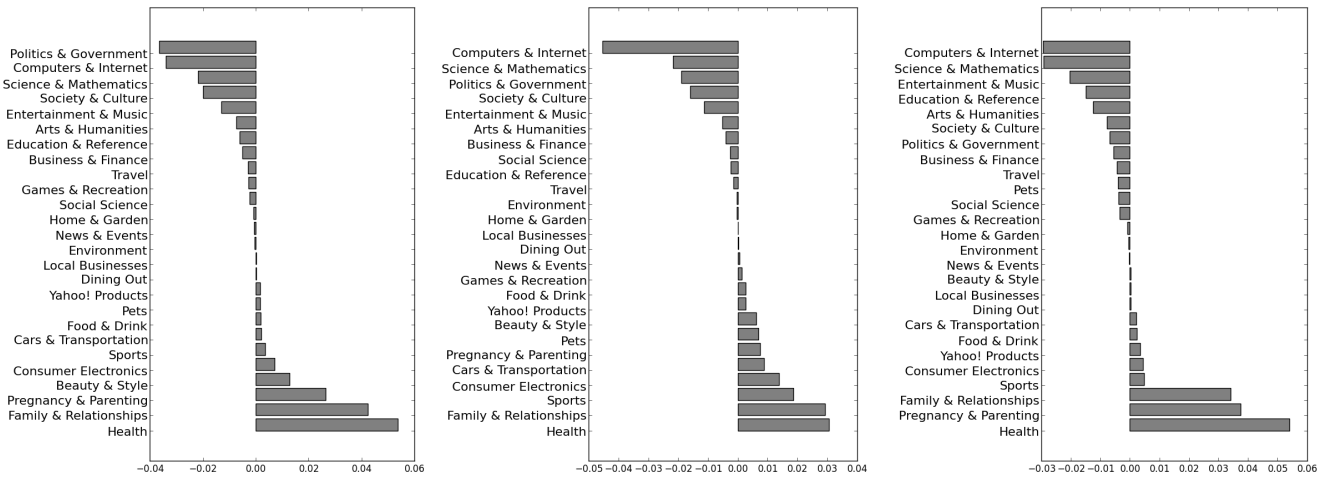
- Pregnancy and birth issues, mostly for females, but also for males (presumably asking about their significant others).
- Health issues, in all groups except males over 50[2].
- Alcohol, for both genders under 20.
- Lesbian, Gay, Bisexual, and Transgendered ("LGBT") issues, for females under 20, and for both genders 25–50.
- Marriage and divorce, for females up to 30, and for males aged 25–30.

These provide additional support to our hypothesis. Note that by conditioning on the demographic, we guarantee that any increase in mobile posting is on top of any increased interest of a particular age and gender group in given topic. For example, the top ranking of "Newborn & Baby" for females under 20 does not necessarily indicate that this is the top category for this age group — that title would be claimed by the "Singles & Dating" category, with Newborn & Baby ranking just 85th out of 1349. Rather, Newborn & Baby is the category with the highest rate of mobile postings, conditioned on age and gender. Overall, of the top-10 categories in this table, the sensitive category ratio is 39% ($31/80$). To appreciate the significance of this observation, consider that most of the thousands of categories in Yahoo! Answers are non-sensitive, ranging in topics from car maintenance to pets.

[2]We omit the over-50 group from the shown results because of data sparsity.

Metric	Desktop Search	Mobile Search	Desktop Questions	Mobile Questions
length in characters	21.50	18.75	Subject: 53.03, body: 343.86	Subject: 46.31, body:300.80
length in words	3.49	3.18	Subject: 12.06, body: 70.82	Subject: 11.07, body: 78.06
characters per word	6.59	6.18	4.30	4.46
percent of male users	54.7	48.4	50.8	51.6
number of posts	$2M$	$588K$	$3.3M$	$784K$

Table 1: Statistics on mobile vs. desktop queries



(a) All users (b) Male users, age 20–24 (c) Female users, age 20–24

Figure 2: Differences in frequencies of question top-level categories, for different populations. A positive value indicates that the frequency of the category is higher for questions posted from mobile devices (and negative means that desktop is higher).

Text analysis

While categories, on average, give a good indication on the topic of a particular question, the diversity between questions in the same category might still be high. To get a better understanding of questions posted from different devices we performed text level analysis.

All questions (subject and body combined) were stemmed using the Porter stemmer. Two language models were built: one for questions posted from mobile devices and another for desktop. Then we sorted all the stems by the difference in probabilities between the mobile and desktop language models. Table 3 (two rightmost columns) shows some terms with the highest differences in probabilities. Results indicate that personal relationship terms like: `girl`, `guy`, `sex`, `talk`, `mom`, etc., have higher probability in questions posted from mobile devices. In addition, questions posted from desktop devices contain the frequent stems `peopl`[3], `find`, `use`, `comput`, `work`, `god`, `obama`, `problem`, `windos`, `game`, `video`, and `file`. This data is in agreement with the results of the question category analysis. Questions posted from desktop devices have a higher probability to contain terms related to god, computers, games, and politics, whereas questions posted from mobile have a higher probability to contain terms related to personal and sensitive information.

Community question answering services cover just a part of users' search needs. Another part of information need is satisfied by search engines. For completeness, we performed an analysis of search queries, as asked from different types of devices. A set of $2.58M$ search queries was extracted from search logs of one of the major search engines. For each query, we know the type of the device from which a question was posted, as well as some user's demographic information (e.g., age and gender).

Here, too, we built two language models — one for each kind of device. Each language model was built after stopword removal, and stemming with the Porter stemmer. Table 3 (two leftmost columns) shows the stems, with the highest differences in probabilities between the mobile and desktop query language models. The terms asked more frequently from mobile devices are adult-themed, while queries from desktops are more frequently navigational and transactional.

But individual question terms might be hard for analysis

[3]This is the stemmed form of "people".

Females

Age	**<20**	**20–25**	**25–30**	**30–50**
N	776,058	446,901	134,474	212,095
1.	**Newborn & Baby**	Newborn & Baby	**Pregnancy**	**Pregnancy**
2.	**Pregnancy**	**Pregnancy**	Newborn & Baby	**Women's Health**
3.	**Pain & Pain Management**	**Pain & Pain Management**	**Trying to Conceive**	Words & Wordplay
4.	Skin Conditions	Skin Conditions	**Women's Health**	Diet & Fitness
5.	Infectious Diseases	**Trying to Conceive**	Baby Names	Makeup
6.	**Marriage & Divorce**	**Women's Health**	**Marriage & Divorce**	**LGBT**
7.	**Beer, Wine & Spirits**	**Marriage & Divorce**	**LGBT**	Newborn & Baby
8.	**LGBT**	Baby Names	Diet & Fitness	Baby Names
9.	**Women's Health**	Diet & Fitness	Friends	Hair
10.	**Trying to Conceive**	Words & Wordplay	Dogs	Singles & Dating

Males

Age	**<20**	**20–25**	**25–30**	**30–50**
N	715,512	559,884	187,581	323,909
1.	PDAs & Handhelds	**Pain & Pain Mgmt.**	Basketball	Rap and Hip-Hop
2.	Yahoo! Search	Yahoo! Search	**Men's Health**	Basketball
3.	Friends	PDAs & Handhelds	**Marriage & Divorce**	**Men's Health**
4.	Football (American)	Basketball	Cell Phones & Plans	Music & Music Players
5.	**Beer, Wine & Spirits**	Skin Conditions	Diet & Fitness	**LGBT**
6.	**Pain & Pain Mgmt.**	Words & Wordplay	Singles & Dating	Xbox
7.	Words & Wordplay	Dental	**LGBT**	Diet & Fitness
8.	Skin Conditions	Football (American)	Military	Words & Wordplay
9.	Cell Phones & Plans	Beer, Wine & Spirits	Words & Wordplay	Cell Phones & Plans
10.	**Pregnancy**	**Pregnancy**	Friends	Singles & Dating

Table 2: Top categories, sorted by rate of posting from a mobile device, stratified by gender and age group
Bold items are potentially sensitive. LGBT stands for "Lesbian, Gay, Bisexual, and Transgendered". Alcohol-related topics are considered sensitive only for the under-20 age group.

and making conclusions. To look at questions' text from a higher level we built an LDA topic model, and then computed frequencies of topics in questions posted from mobile and desktop devices. The GibbsLDA++ implementation (Phan and Nguyen 2007) was used, with the number of topics set to 50. Some of the topics, estimated by the model are: pregnant (high frequency words in this topic include: day, period, pregnant, test, week, sex), internet slang (u, lol, ok, thx, plz, btw), beauty (hair, look, color, skin, face, nail), politics (obama, romney, support, us, tax, govern), Internet (facebook, google, instragram, email, send, post), computer games (game, compute, download, play, laptop, connect, xbox), and money (buy, money, pay, cost, sell, card).

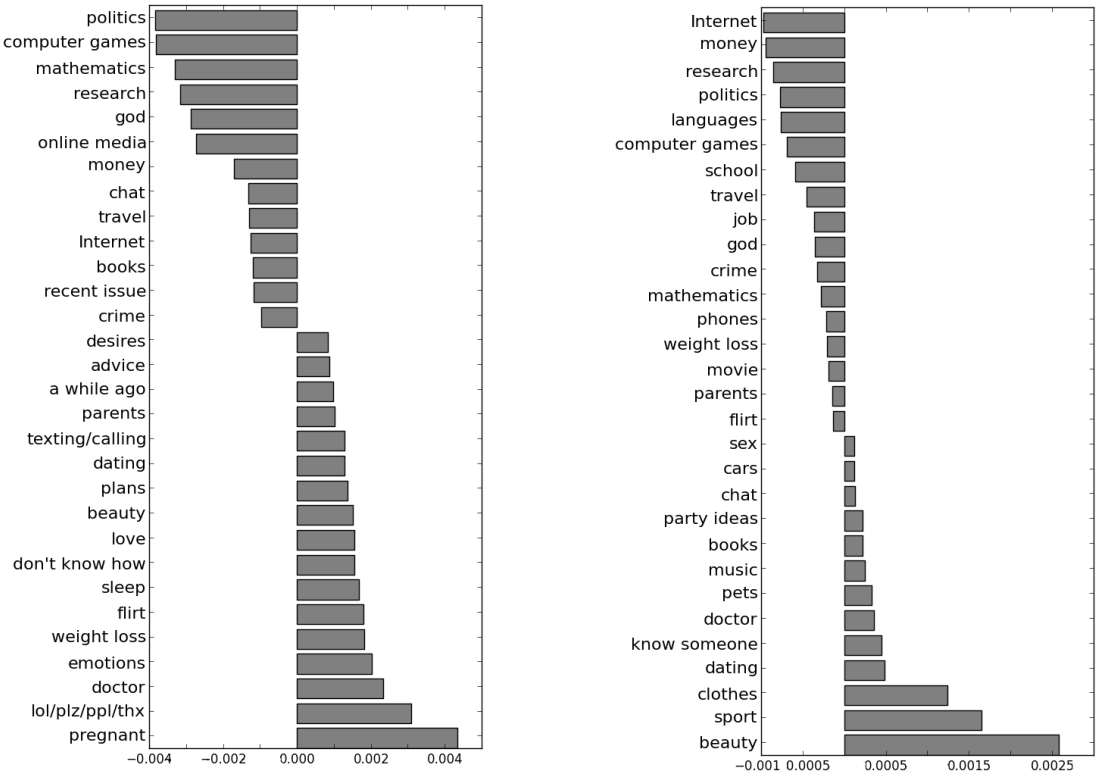
As in the previous experiments, we computed differences between frequencies of LDA topics in questions posted from mobile and desktop devices. Figure 3a presents the topics sorted by this difference. Positive values mean that the topic occurs more frequently in the questions posted from mobile devices and negative — from desktop devices.

Again, results of the LDA analysis are in agreement with both category and text analysis. Topics related to sensitive questions, personal relationships, health issues, etc. are more frequently asked from mobile devices, and politics, computers, games, religion, science, etc. are asked more from desktop devices. We further split users by age and gender and looked at data conditioned on the demographics. The results (not shown) support the above-mentioned conclusions, the only difference is a shift of interest for people of different age and gender, which is similar to those shown for question categories.

To look at the search queries from the topical level, again we used the LDA model, estimated on questions, and inferred the distribution of topics in queries asked from desktop and mobile devices. Figure 3b shows the plot, where topics are sorted by the difference of frequencies in queries asked from mobile and desktop devices. As in earlier plots positive values mean that a topic has higher frequency to appear in a query issued from a mobile device.

The results agree with the previous analysis of questions posted to a CQA website. For example, the mostly mobile topics are beauty, sports, dating, doctor, and a topic we called "know someone", where the posts ask if other users know someone who can help, give advice, etc. On the other hand, the mostly desktop categories are Internet, finance, research, and politics. This generally supports a positive answer to the research question RQ1b. The sports topic stands out from the list, but its popularity on mobile devices could be well explained: users might check on latest sport competition results from their phones.

Queries asked by users of some specific age and gender have the same properties as in the aggregated case described above. Of course, some topics have higher popularity for



(a) CQA questions

(b) Search queries

Figure 3: LDA topics sorted by difference of frequencies in questions posted from mobile and desktop devices The chart shows the topics with the most extreme values (i.e., the middle of the 50-topic list is omitted).

people of a specific age or gender. For example, politics tend to be less popular among youngsters, and asked more about by older people. But in general, queries issued from mobile devices are about personal matters more often then when issued from desktops.

Location and Mobility Analysis

To explore the location issue, we collected web view ("click") data from question askers in our data set, during the same period (in July and August 2012). Each page view was geo-located from its IP address using an internal product, and used at the resolution of zip codes. To emphasize, the geo-tag is not accurate to the user's exact location (as it would be, if the information originated from the device's GPS sensor). It is coarser, with the accuracy varying depending on many technical factors, but typically at the level of several city blocks.

We then partitioned the data by user, and sorted by the timestamp. This gives us the "itinerary" of the user, during the five weeks of the experiment, in terms of zip codes where he or she accessed the web. As an example, assume the user Alice visited the following locations, in order: A, A, A, A, B, A, A, A, and the user Charlie exhibited the sequence C, D, C, D, C, D, C. Then we compute, for each user, the most frequent location — in our example A for Alice and C for Charlie. Additionally, we compute the number of runs (changes from the previous value) in the sequence — three for Alice and seven for Charlie. This will capture the notion that Alice is mostly stationary when accessing the web, whereas Charlie regularly commutes between his web access location.

In our data, 60.2% of users access the web from only one location, 80.0% do so from up to two locations, and 99.6% do so from up to nine locations. In addition, 60.2% of users have one run (by definition), 66.9% have up to two runs, and 89.3% have up to nine runs.

Next, to each question in the data set (which is a special case of a page view), we applied the same geo-location logic, and tagged it with the per-user rank of the popularity of its location. Going back to the example, in Alice's sequence, all accesses which match question postings from A would be given the rank 1, and B would rank as 2. In particular, this classifies questions into those that were posted from the user's primary location (rank equals one), and others (rank greater than one).

We also filtered the data to consider only the users with at least four runs, to exclude the ones which hardly move around. We also removed users with more than 20 different locations, as they are most likely constantly on the move. Figure 4 shows the top-level categories, sorted by the rate of posting to them from the non-primary, non-desktop location. The least-frequent category by this metric is Ya-

Desktop Search	Mobile Search	Desktop Questions	Mobile Questions
us	tattoo	peopl	like
facebook	porn	find	help
youtube	photo	use	girl
online	hot	comput	like
download	bikini	work	feel
job	wed	god	friend
citi	funni	obama	period
free	pictur	problem	eat
site	hair	window	mom
game	style	video	pleas
sale	girl	file	guy
school	imag	program	sex
price	nail	link	talk
service	love	game	pregnant

Table 3: Top stems by difference of probabilities between the mobile and desktop question language models

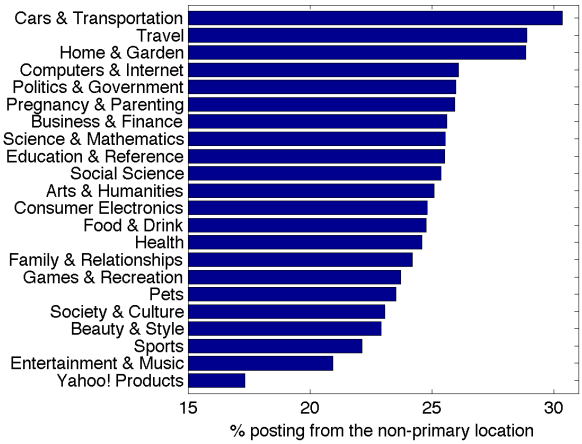


Figure 4: Rates of posting from the non-primary location while using a non-desktop device

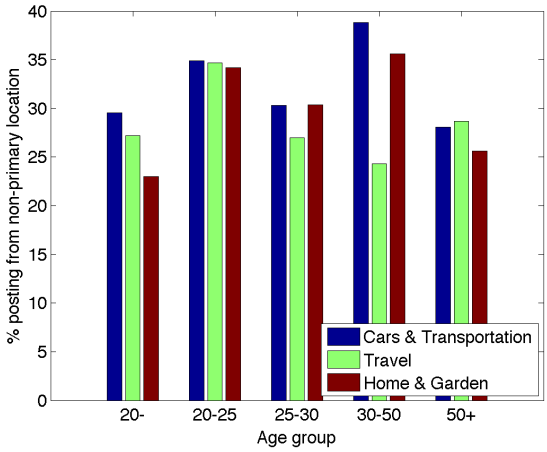


Figure 5: Rates of posting from the non-primary location while using a non-desktop device, by age group

hoo! Products. The top categories, with rates of 28% to 30%, are Cars & Transportation, Travel, and Home & Garden. While the Travel category is understandable, the others are not as clear. Cars & Transportation has sub-categories for aircraft, boating, and so on, as well as an active subhierarchy of car makes, where repairs and purchase decisions are discussed (typical example: "Will 28 inch rims look good on a Chrysler 300?"). Home & Garden has categories for decorating, do-it-yourself, and maintenance (typical example: "How can i hang shelves without putting holes in the wall?"). We break usage for these top three categories in Figure 5. We see that the 20–25 group has consistently high rates of posting from the infrequent location, and that the 30–50 group has very high rates for Cars & Transportation and Home & Garden (but not for Travel). It's possible that the propensity of younger users to post from varied locations is higher, but that the very young (under 20) don't move around that much — presumably just commuting between home and school. As age increases, vehicle and house ownership rates grow, and drive increased interest in the respective categories. We propose these hypotheses without proof, and while we feel there is interest in exploring this particular need further, it is outside the scope of the current work.

Physical Location and Sensitive Posting

Since we also kept track of the device type indication discussed above, each question in our resulting data set had two independent features:

- Question was posted from mobile or desktop device.
- Question was posted from primary or non-primary location for the user.

To control the level of sensitivity, we identified three types of questions, in increasing order of sensitivity:

1. Questions from the popular "Polls & Surveys" category ($n = 57621$). Generally those map to idle chat without any sensitive content, and are used as a control group. Typical example: "*Can you hear your heart beat?*". Marked as "not sensitive" below.
2. Questions from the Pregnancy category ($n = 9812$). These might contain sensitive material, as discussed above. But not necessarily — they could just as well be about pregnant pets owned by the asker. Typical example: "*19 weeks pregnant. Tingling legs won't go away?*". Marked as "potentially sensitive" below.
3. Questions about fear or uncertainty of a possible and unplanned pregnancy ($n = 1542$). These are from the Pregnancy category used above, and then matched against a hand-written list of regular expressions. Typical example: "*16 and I think I'm pregnant?*". Marked as "sensitive" below.

We stratified the questions by age group and gender, as before, to discount the effect of these variables on mobile posting rates. As a first observation, we find that the more sensitive a question is, the higher the rate of it being posted

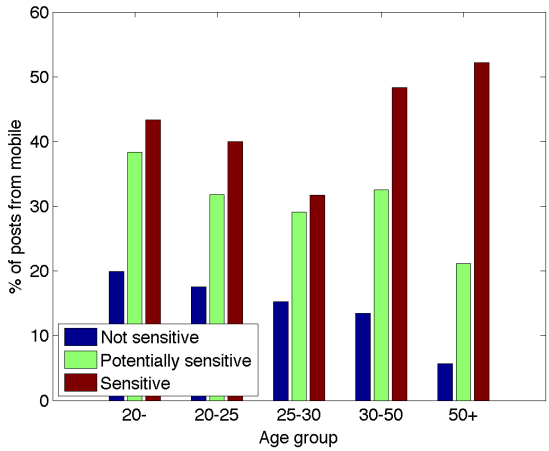


Figure 6: Rates of posting from a mobile device, by question type and age group

from a mobile device. See Figure 6. Results are significant (at $p < 0.01$), with the exception of the age groups 25–30 and 50+, between potentially sensitive and sensitive. In addition, we checked if posting from a mobile device is correlated with posting from an infrequent location. We found that for the under-20 age group, they are correlated for both the sensitive and the potentially sensitive posts (χ^2 test for independence, $p < 10^{-3}$, $n = 4240, 812$, respectively). In other words, if a question is posted from a mobile device, then it is more likely to also be posted from an infrequent location, on top of the marginal probability of seeing it from an infrequent location. In addition, for the non-sensitive posts, such correlation does not exist for the under-20 group ($p = 0.66, n = 22396$).

These results, framed in terms of research question RQ2, show that when posting a sensitive question from a mobile device, the asker is also more likely to be in an unique location. For non-sensitive posting, there is no such correlation.

Discussion

We use personal details self-reported by the user, including age, gender, and zip code, and those could be biased. However, self-reported gender was previously found to be consistent with other contributed data in 96% of cases (Pelleg, Yom-Tov, and Maarek 2012). As for zip code, we checked the self-reported value against the zip code of the most frequent location of access. For 70% of users, the distance is less than 70 miles, indicating that this piece of information is generally accurate. Self-reported age is documented to be over-reported for teenagers on social web sites (Boyd et al. 2011). The effects on our statistics is minimal, since we lump everyone under 20 into a single group.

The term "mobile" might mean different things, depending on context. While some associate it with cellular data access, as much as 37% of this traffic comes over Wi-Fi (ComScore 2011). Note that many cellular-capable devices prefer a Wi-Fi connection, and will use an available hot spot even if they contracted a data plan. We have also observed a significant volume of content contributed from iPod devices, which technically are not smartphones (but have similar capabilites in terms of web interaction, when used over Wi-Fi). Use of VPNs and proxies might also affect the data, as well as usage of a "personal hotspot" feature which would allow a laptop computer to connect over a cellular data link. Our analysis is generally immune to these effects as we use a browser-provided string.

Mobile devices might be shared — for examples when children borrow their parents' smartphones or tablets to play games. This could distort our results, but we feel this use case is unlikely to result in contributed content and therefore would not affect our data. In addition, we exclude all data from tablet computers, which are more frequently shared.

The resolution of our location information is quite coarse, at the level of a zip code. Ideally we would like to identify it as classroom, coffeeshop, bus, or even a supermarket checkout line. This seems beyond the limits of current technology. More realistically, it might be possible to classify our locations as "home" or "work", but the required algorithmics are outside of scope.

Despite our efforts to control for age by stratification, the resolution might be too coarse, and allow age effects to slip in. Moreover, we do not control for income, which is also related with mobile usage in intricate ways. We only used data where the user logged into the service. For Yahoo! Answers this is benign, but for web view data it could introduce a bias, as the rate of users who log in is generally low on mobile devices.

Conclusion

Smartphones and other web-capable mobile devices are becoming ubiquitous. While these devices are still improving, they are already overtaking Desktop PCs in popularity. One of the most important reasons for this, is personalization: the user's personal address book, web credentials, and search history are always on the person. Thus the modern mobile devices are truly becoming a personal digital assistant — companions, that will store our personal data, including our queries, questions, and other information needs, and are expected to keep it in confidence. Take for example sexting — sending or receiving sexually suggestive self-photos, which is practiced by at least 30% of 17-year-olds (Lenhart 2009). While this kind of activity could technically be performed on a PC, equipped with a camera, it would have never become so popular if it weren't for mobile devices. Therefore the need arises to examine what aspects of online behavior are affected by this new mode of computing, and in particular the expression of sensitive information needs.

This work breaks new ground in a number of ways, not the least in that it parallels early work analyzing information needs of users of web search engines (Broder 2002; Spink et al. 2002; Rose and Levinson 2004) but at much larger scale and diversity of users. It focuses on *social* information seeking as opposed to automated web search. Questions submitted to CQA services tend to be much more detailed and rich than web search queries, allowing us to perform deeper analysis of the stated information needs and the way the questions tend to be expressed to others. As a result,

we confirm (and disprove) previous intuitions about mobile information seeking, and present new findings and insights based on solid empirical evidence.

Specifically, we first showed that both search queries and CQA questions are more sensitive and personal when issued from a mobile device. Most prominent topics include personal health, sexual orientation, and relationships. Second, we show that location is positively correlated with postings of sensitive questions: posting from a mobile device is associated with questions about unplanned pregnancies, and doing so from a non-frequent location is more prevalent than one would expect. Finally, our study uses the largest dataset by far of comparable studies (with millions of questions, queries, and users). This allows us to control for age and gender, and to produce usage statistics at unique levels of detail. In the future, our data could support several follow-up projects, including studying the mobile information usage by middle-aged and senior citizens, characterizing tablet computer use, and multi-party information needs.

Acknowledgements

We wish to thank Yoelle Maarek for helpful comments and suggestions. This work was made possible by the Yahoo! Faculty Research Engagement Program.

References

Ahern, S.; Eckles, D.; Good, N. S.; King, S.; Naaman, M.; and Nair, R. 2007. Over-exposed?: privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '07, 357–366. New York, NY, USA: ACM.

Aoki, K., and Downes, E. J. 2003. An analysis of young people's use of and attitudes toward cell phones. *Telematics and Informatics* 20(4):349–364.

Boyd, D.; Hargittai, E.; Schultz, J.; and Palfrey, J. 2011. Why parents help their children lie to facebook about age: Unintended consequences of the 'children's online privacy protection act'. *First Monday* 16(11).

Broder, A. 2002. A taxonomy of web search. *SIGIR Forum* 36(2):3–10.

Church, K., and Smyth, B. 2009. Understanding the intent behind mobile information needs. In *Proceedings of the 14th international conference on Intelligent user interfaces*, IUI '09, 247–256. New York, NY, USA: ACM.

ComScore. 2011. Smartphones and tablets drive nearly 7 percent of total u.s. digital traffic. Press Release. `http://www.comscore.com/Insights/Press_Releases/2011/10/Smartphones_and_Tablets_Drive_Nearly_7_Percent_of_Total_U.S._Digital_Traffic`.

Cooper, G. 2001. The mutable mobile: social theory in the wireless world. In *Wireless world*, 19–31. Springer-Verlag New York, Inc.

Ghose, A.; Goldfarb, A.; and Han, S. P. 2010. How is the mobile internet different? *Social Science Research Network Working Paper Series*.

Hasler, L., and Ruthven, I. 2011. Escaping information poverty through internet newsgroups. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Karlson, A.; Meyers, B.; Jacobs, A.; Johns, P.; and Kane, S. 2009. Working overtime: Patterns of smartphone and PC usage in the day of an information worker. In Tokuda, H.; Beigl, M.; Friday, A.; Brush; and Tobe, Y., eds., *Pervasive Computing*, volume 5538 of *Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer Berlin Heidelberg. chapter 27, 398–405.

Kıcıman, E. 2012. OMG, I have to tweet that! a study of factors that influence tweet rates. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*.

Lee, U.; Kang, H.; Yi, E.; Yi, M.; and Kantola, J. 2012. Understanding mobile Q&A usage: an exploratory study. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, 3215–3224. New York, NY, USA: ACM.

Lenhart, A. 2009. Teens and sexting. `http://www.pewinternet.org/Reports/2009/Teens-and-Sexting.aspx`.

Mancini, C.; Thomas, K.; Rogers, Y.; Price, B. A.; Jedrzejczyk, L.; Bandara, A. K.; Joinson, A. N.; and Nuseibeh, B. 2009. From spaces to places: emerging contexts in mobile privacy. In *Proceedings of the 11th international conference on Ubiquitous computing*, Ubicomp '09, 1–10. New York, NY, USA: ACM.

Palen, L.; Salzman, M.; and Youngs, E. 2000. Going wireless: behavior & practice of new mobile phone users. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 201–210. ACM.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In Adamic, L. A.; Baeza-Yates, R. A.; and Counts, S., eds., *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Pelleg, D.; Yom-Tov, E.; and Maarek, Y. 2012. Can you believe an anonymous contributor? on truthfulness in Yahoo! answers. In *2012 International Confernece on Social Computing (SocialCom)*.

Phan, X.-H., and Nguyen, C.-T. 2007. GibbsLDA++: A c/c++ implementation of latent dirichlet allocation (LDA).

Rose, D. E., and Levinson, D. 2004. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, 13–19. New York, NY, USA: ACM.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, WWW '10, 851–860. New York, NY, USA: ACM.

Spink, A.; Jansen, B. J.; Wolfram, D.; and Saracevic, T. 2002. From E-sex to E-commerce: Web search changes. *COMPUTER* 35(3):107–109.

Teevan, J.; Karlson, A.; Amini, S.; Brush, A. J. B.; and Krumm, J. 2011. Understanding the importance of location, time, and people in mobile local search behavior. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, 77–80. New York, NY, USA: ACM.

Troianovski, A. 2013. The web-deprived study at McDonald's. The Wall Street Journal.

Wang, Y.; Huang, X.; and White, R. 2013. Characterizing and supporting Cross-Device search tasks. In *6th Annual International ACM WSDM Conference on Web Search and Data Mining (WSDM 2013)*.

Wei, R., and Leung, L. 1999. Blurring public and private behaviors in public space: policy challenges in the use and improper use of the cell phone. *Telematics and Informatics* 16(1–2):11–26.

Yi, J.; Maghoul, F.; and Pedersen, J. 2008. Deciphering mobile search patterns: a study of Yahoo! mobile search queries. In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, 257–266. New York, NY, USA: ACM.