Aiming at a More Cost-Efficient Census Via Online Data Collection: Privacy Trade-Offs of Geo-Location

PRELIMINARY DRAFT - Research in progress

Laura Brandimarte
Alessandro Acquisti

Carnegie Mellon University

NCRN Meeting
May 22, 2014
Census privacy trade-offs

• In an effort to reduce costs, the 2020 Census will take place via the classical paper questionnaire and the online questionnaire

• The success of the online questionnaire may depend on
  1. providing adequate incentives for a sizable portion of respondents to transition from the offline to the online format
  2. addressing individuals’ security and privacy concerns
As Internet responses increase, proportion of cases without ID may increase too

• Entering a pre-assigned ID onto an Internet form is not easy...
Possible solution:
Pre-fill online forms with location data

• How can the Census gather accurate location data?
  – Obtain the best self-reported address...
  – Use location-based services via mobile devices...
  – Ask respondents to confirm/correct their location on a map...
Research questions

1. What is the effect of showing location data (making respondent aware of location identification) on willingness to provide further personal information?
   – Will this provide an incentive to complete the questionnaire, thus increasing response rate?
   – Will it raise privacy concerns, and thus backfire?

2. Are there location privacy concerns specific to the Government, or is the Government (and, specifically, the Census Bureau) trusted?
Related literature

• Effects of general privacy concerns on online behavior (Gideon, Cranor, Egelman & Acquisti, 2006; Tsai, Egelman, Cranor & Acquisti, 2011; White, Zahay, Thorbjørnsen & Shavitt, 2008)

• Ubiquitous computing: privacy concerns related to disclosure of location information (Agre & Rotenberg, 1997; Barkhuus & Dey, 2003; Doheny-Farina, 1994; Garfinkel, 2001)

• Effects of privacy concerns on adoption of location-aware technologies (Zweig & Webster, 2002)

• Theoretical models to identify determinants of privacy concerns due to location tracking (Junglas & Spitzmuller, 2005; Aloudat, Michael, Chen & Al-Debei, 2013)

• Empirical investigations of location privacy concerns (Sadeh et al., 2009; Toch et al., 2010)
Methodology

• Six between-subject randomized experiments (we only cover a subset of those in today’s presentation)
  – Experiments 1A and 1B
  – Experiments 2, 3, 4, and 5

• Manipulations:
  – Geo-location awareness
  – Institution requesting data
  – Sensitivity of questions

• Dependent variables
  – Perceived intrusiveness of questions
  – Propensity to provide sensitive information
    • Engagement in unethical behaviors (Brandimarte, Acquisti & Loewenstein, 2013; Joinson, Woodley, & Reips, 2007; Phelps et al. 2000; Weisband & Kiesler, 1996)
    • Census-related questions (demographics and living arrangements)

• Mediator: privacy concerns
Experiments 1A and 1B

• Three conditions (we captured location in all of them):
  – Control
  – Geo-Located (Country, State, City, Zip code)
  – Requested Location (Country, State, City, Zip code)
• 7 Census-related questions
• 16 sensitive behavior questions
  – Why did we use them?
• Exit questions about privacy concerns, feeling tracked or monitored
• Different samples: MTurk in Experiment 1A, CMU CBDR in Experiment 1B
Experiment 1A - Results

- 403 Mturk workers (37% female, $M_{age} = 29.8$, $SD = 9.4$)
- Census-related score: ceiling effect
Experiment 1A - Results

- Sensitive behaviors, average disclosure score:
Experiment 1A - Results

- Sensitive behaviors questions: Panel specification, probit estimation

Table 1. Marginal effects for panel probit estimation – Experiment 1A. Standard errors in brackets. ** indicates significance at the 1% level; * at 5% level.

<table>
<thead>
<tr>
<th></th>
<th>Disclosure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>.138*</td>
</tr>
<tr>
<td></td>
<td>(.070)</td>
</tr>
<tr>
<td>Requested_Location</td>
<td>.006</td>
</tr>
<tr>
<td></td>
<td>(.070)</td>
</tr>
<tr>
<td>Male</td>
<td>.194**</td>
</tr>
<tr>
<td></td>
<td>(.061)</td>
</tr>
<tr>
<td>Age</td>
<td>-.003</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
</tr>
<tr>
<td>White</td>
<td>.213**</td>
</tr>
<tr>
<td></td>
<td>(.065)</td>
</tr>
</tbody>
</table>

Number of groups = 403
Wald χ² (5) = 26.78
Experiment 1B

- Same as 1A, but different population
- 186 participants from CMU CBDR pool (60% female, $M_{\text{age}} = 26.8$, $SD = 10.8$)
- Consistent results (but larger power)
Experiment 1B

- **Geo-Located condition**, sensitive behavior questions:
  - No difference in the Requested Location
  - If geo-location is successful, people do respond by revealing less, but if the procedure is not accurate, then the manipulation loses effectiveness
Limitations

• Why did we not observe a significant difference between the two treatments?
  – In the Requested Location condition, we only got the zip code exactly right in 25 out of 134 cases (assuming self-reports are ground truth)
  – For approximately half of our sample, only the first two digits of the captured zip code were correct, which may have decreased the feeling of being tracked, reducing privacy concerns

• Possible solutions:
  – Hide the last two digits of the zip code (ongoing)
  – Use a better location tracking tool and only analyze the data for which we are able to capture the correct geo-location
Limitations

• Why did we observe a ceiling effect for Census-related questions?
  – Perceived low sensitivity of these questions (no PII could be collected)
  – How can we “make” such questions sensitive? Perhaps by manipulating who requests them?
  – See Experiment 2 and Experiment 3
Experiment 2

• Attempt to explain ceiling effect and find a way around it
• 4 conditions, differing by the entity requesting the data:
  – Researchers
  – U.S. Government
  – Census Bureau specifically
  – Advertisers
• 402 Mturk workers evaluated the level of intrusiveness of 9 of the sensitive behaviors questions used in the previous experiments, all the Census-related questions therein, and other PII requested in the 2010 Census form
Experiment 2 – Results

- Census-related questions: $M_{\text{intrusive}} = 3.12$, $SD = 1.31$
- Sensitive behaviors questions: $M_{\text{intrusive}} = 4.74$, $SD = 1.66$, paired $t(401) = -15.971$, $p < .001$
Experiment 2 – caveat

- Mturk ToS do not allow researchers to request PII, which may have caused MTurk workers to rate PII as more sensitive if requested by researchers or advertisers than by governmental institutions.
Experiment 3

- N = 694 MTurk Workers (41.2% female, $M_{age} = 31.1$, $SD = 10.6$)
- 3x3 between-subjects, manipulating geo-location (as in Experiments 1A-B) and Institution (Researchers, Census Bureau, Government, as in Experiment 2)
- 12 Census-related questions
- 16 sensitive behavior questions
- Exit questions about privacy concerns, feeling tracked or monitored
Experiment 3 - Results

- Census questions:
  - Again ceiling effect

- Sensitive questions:
  - Main effect of geo-location
    \((F(2,682) = 4.165, p < .05)\)
    - Higher disclosure rates in the Control conditions that in the two geo-location conditions \((t(685) = 3.22, p = .001)\)
    - Requested Location and the Geo-Located condition did not differ from each other \((t(685) = .48, p > .10)\)
    - Effect is strongest for Government institutions (less trusted?)
  - Main effect of type of institution requesting the data \((F(2,682) = 4.493, p < .05)\)
    - Higher disclosure rates if Researchers requested info as compared to Census or Government \((t(685) = 3.40, p = .001)\)
  - No significant interaction

Covariates appearing in the model are evaluated at the following values: Gender = 1.4121, Age = 31.1311, White = .7666
Experiment 4

- Explaining lower trust (higher concerns for geo-location) in Government institutions found in Experiment 3
- Effect of surveillance primes and entity requesting data on perceived intrusiveness of questions
- In light of recent developments regarding government data collection and surveillance programs, we expect perceived intrusiveness of Census-related questions to be higher for governmental institutions if surveillance primes are present
Experiment 4

• N = 603
• Design: 3x2 between-subject
  – Manipulate the alleged entity requesting the data (Governmental institution, Census specifically, Researchers) and the presence of surveillance priming (participants solve anagram either containing or not containing a surveillance prime)
• DV: same as Experiment 2 (perceived intrusiveness of Census-related questions and sensitive behaviors questions)
Experiment 4 – Results

<table>
<thead>
<tr>
<th>Perceived Intrusiveness</th>
<th>Without Prime</th>
<th>With Prime</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Census</strong></td>
<td>3.5 (±0.5)</td>
<td>3.2 (±0.4)</td>
</tr>
<tr>
<td><strong>Sensitive</strong></td>
<td>4.8 (±0.6)</td>
<td>5.1 (±0.7)</td>
</tr>
</tbody>
</table>

- Census questions less intrusive than sensitive behaviors questions (p < .001)
- Census questions: main effect of institution (F(2, 596) = 5.476, p < .01), but no effect of priming. No significant interaction
- Sensitive questions: main effect of institution (F(2, 596) = 15.721, p < .001) and priming (F(2, 596) = 4.327, p < .05). No significant interaction
Conclusions

• Our results (always ceiling effect on Census questions) seem to suggest that pre-population of location data in the Census form will not increase privacy concerns, and thus may successfully incentivize citizens to fill the form online...

• ...but awareness of geo-location decreases people’s willingness to provide sensitive information...

• ...and people seem less comfortable disclosing to Census or Government institutions than to researchers

• Potential problem if actual Census forms were perceived as privacy intrusive
Conclusions

• This could extend to other (non-location) data (e.g., administrative data from DMV)

• Alternatives to geo-location:
  – Ad campaigns focusing on the completion of the form as a duty
  – Emphasize difference between geo-location and location tracking
Thank you

lbrandim@andrew.cmu.edu
acquisti@andrew.cmu.edu