Toward Harmonizing Self-reported and Logged Social Data for Understanding Human Behavior

Conference Paper • January 2017

2 authors, including:

Vivek K. Singh
Rutgers, The State University of New Jersey

Some of the authors of this publication are also working on these related projects:

- Cyberbullying detection and prevention View project
- Situation Recognition View project
Toward Harmonizing Self-reported and Logged Social Data for Understanding Human Behavior

Vivek K. Singh  
School of Communication and Information, Rutgers University, NJ, USA.

Arushi Jain  
Department of Economics, Indian Institute of Technology Kanpur, India.

ABSTRACT
While self-reporting remains the most common method to understand human behavior, recent advances in social networks, mobile technologies, and other computer-mediated communication technologies are allowing researchers to obtain detailed logs of human behavior with ease. While the logged data is very useful (and accurate) at capturing the structure of the user’s social network, the self-reported data provides an insight into the user’s cognitive map of her social network. Based on a field study involving 47 users for a period of ten weeks we report that combining the two sets of data (self-reported and logged) gives higher predictive power than using either one of them individually. Further, the difference between the two types of values captures the level of dissonance between a user’s actual and perceived social behavior and is found to be an important predictor of the person’s social outcomes including social capital, social support and trust.

Author Keywords
Self-reported; Call-log data; Bias; Dissonance Coefficient; Socio-Mobile Behavior; Social Ties

ACM Classification Keywords
H.1.2 User/Machine Systems, Human Factors

INTRODUCTION
Tracking human behavior is important for multiple HCI domains including eliciting user experiences, behavior change, and computational social science [18, 19, 25]. While the most common method for eliciting human behavior remains self-reports, multiple researchers have found issues (inaccuracies, biases, adherence) with self-reports and suggested ways to counter them (e.g. [3, 18, 25]). One trend in recent years has been the use of logged sensor (e.g. phone) data to capture user behavior for longitudinal studies, and such automated logged methods are often touted to be superior replacement for the self-reported data [3, 15, 22]. For example, [15] argues a case of using accurate phone use data, as incorrect data result in false observations and unfounded associations between different social behaviors and outcomes of interest (e.g. associations between number of phone calls and civic engagement).

While accuracy is important, we posit that the common practice of considering one type of data to be "better" than the other may be counterproductive because the two methods (self-reported and logged) capture inherently different dimensions of human behavior. While the logged data is very useful (and accurate) at capturing the structure of the user’s social network, the self-reported data provides an insight into the user’s cognitive map of her social network. Indeed many of the social outcomes of interest (e.g. social capital, social support) considered in such studies are just as much dependent on a user’s cognitive map or the perception of social processes as they are connected to the actual structure of the network. Pushing this line of reasoning further, we posit that the (dis)similarity between the two types of data, structural and cognitive – defined as a dissonance coefficient here - can also be a useful predictor of individual outcomes of interest (e.g. social support).

It is well-known that individuals create their own “subjective social reality” from their perception of the world [8, 9] and an individual's construction of social reality, not the objective input, may dictate their behavior, and hence outcomes, in the social world. Thus, such biases may sometimes lead to perceptual distortion, or inaccurate judgment. While measuring such a distortion in itself has been challenging in the past, the datasets containing both self-reported and logged data provide a natural opportunity to quantify this distortion.

An important reason for collection of self-reported data in longitudinal studies is to identify its associations with social outcome variables of interest. For example, to study the interconnections between phone use and civic engagement [15]. The significance of the associations found as well as their effect sizes can have significant scholarly and policy implications. Hence, understanding the effects of using different types of data (logged, self-reported, their difference, and the combination) on the associations found can be of vital practical importance. Based on a ten week study involving both self-reported and logged information...
about phone-based social interactions involving 47 participants, we report that (1) using a combination of logged and self-reported data yields more predictive power at outcomes of interest (e.g. social capital) than considering either of them in silos; and (2) dissonance coefficient i.e. the dissimilarity in the self-reported and logged data, is in itself an important predictor of different outcomes of interest including social capital, social support and trust.

**BACKGROUND AND RELATED WORK**
The reliability of measures for human relationships has been the subject of sharp debate over few decades now. Over the last 30 years, there have been studies which analyzed the accuracy of such information and have concluded that people are imprecise in reporting about their own social network patterns and misrepresent who they talk to [7, 12, 17].

When researchers collect data on social networks, they ask people to recall their interactions but this can be problematic as individuals may forget relevant persons in response to network elicitation questions [17, 20]. The evidence of forgetting or low degrees of recall in response to prompts and cues further suggest such forgetting [10, 28]. When Bernard, Killworth, & Sailer analyzed triads (groups of 3 people) [12] and cliques (closed knit-groups) [13] in their experiments, they found out that the cognitive data reported by respondents and behavioral data had virtually no agreement between them if studied triad by triad. However, the data display the same structure if aggregated at a coarser granularity.

The recent advancement of mobile communication and technology have allowed researchers to obtain accurate logged data. Multiple efforts have compared self-reported and logged data [3, 15]. These studies have reported general consistencies but specific misalignments in values such as the frequency of mobile phone use [3, 4, 15]. While “forgetting” remains the most common explanation for the discrepancy between self-reported and logged data, it is by no means the only explanation. Multiple studies have reported different biases including attractiveness bias, sociability bias, and expansiveness bias as possible explanations and hence also considered different factors such as network size, salience, behavioral specificity and tie-strength tied to interpret the differences found [4, 5, 6].

**STUDY**
The data used in this paper was obtained as part of the Rutgers Well-being Study. Participants in this study were asked to attend three in-person sessions for surveys and install a mobile application onto their smartphone which recorded anonymized call meta-data (calls made or received: number, times and anonymized id but no actual audio). The app also recorded SMS and GPS meta-data, which is not relevant to the current discussion, though.

The study included 59 participants. However, some of the participants did not complete all the surveys, and some did not enter their unique identifying code consistently across different surveys, resulting in a set of 47 (31 male, 16 female) participants for whom we have the mobile-based data as well as the scores for the five surveys of interest (more details on surveys presented later). The survey order was randomized for different participants. Participation in this study was voluntary to the study was incentivized monetarily. The participants were compensated up to a sum of $100 on successful completion of the study over the ten week period.

The participants were also asked other questions about their demography (age, gender, marital status, level of education, and family income level), and their social activities. The most common age group was 18-21 years, the most common education level was “some college” and the median annual family income was in the range $35,000-$49,000 and 92% of them were single. All personnel involved with the study underwent human subject training and IRB certification. All data were anonymized before analysis.

**MEASURES**

**Social Capital**
This survey was based on the Williams (2006) Internet Social Capital Scales (ISCS) [32]. While the original survey was intended to measure ‘online’ social capital and contrast it with the ‘offline’ social capital, we were interested in a general purpose interpretation of social capital. Hence, we removed the words ‘online’ and ‘offline’ in the survey to get a general purpose understanding of a user’s social capital. For example the question “when I feel lonely, there are several people online (respectively offline) I can talk to”, was simply replaced by “when I feel lonely, there are several people I can talk to.” The Social Capital score was computed by adding the scores for all the relevant questions (with reverse adjustments).

**Social Support**
This survey consisted of 18 questions (related to the availability of companionship, assistance, or other types of support when one needs it) based on Sherbourne & Stewart’s Modified Social Support Survey [24]. The participants were asked to answer the question on a 5-point scale ranging from: None of the Time (1) to All of the Time (5).

**Trust**
This survey uses general statements to measure participants’ beliefs about honesty and trustworthiness of others, in general. The questions came from the Yamagishi and Yamagishi’s General Trust Scale [29]. This survey consisted of 6 questions (related to the beliefs of an individual regarding honesty) and the participants were asked to answer the question on a 5-point scale ranging from Strongly Disagree (1) to Strongly Agree (5). Trust score was the percentage of the maximum score possible.

**Self-report of social behavior**
The participants were also asked questions about their phone-based social behavior. This included questions about the number of calls made, number of people called, and the frequency of calling the top three contacts. (Details follow).
**Smartphone data measures**

The 47 users in the dataset made a total of 22,573 calls during the 10 week period. This corresponds to roughly 52 calls per user per week. A descriptive summary of the survey data for the same participants is shown in Table 1.

<table>
<thead>
<tr>
<th>Data type</th>
<th>Data points (54 participants)</th>
<th>Mean (each participant)</th>
<th>Median (each participant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calls</td>
<td>24,573</td>
<td>523</td>
<td>312</td>
</tr>
</tbody>
</table>

Table 1. Summary of call data considered in this study.

**Characterizing Social Processes**

Based on a survey of existing literature we have identified a set of features to represent a user’s social behavior. These features were inspired by the literature on comparisons of call-log data and self-reported data [3, 18] and communication patterns in the social networks [21]. Specifically, prior literature in the area has identified the role played by network size (number of calls, number of people called), strength of ties (most frequent contact vs. less frequent contact), and time elapsed on the quality of the interaction best represented in the data. 

We used the LASSO regression technique as we had limited number of data points. LASSO adds penalty for each additional feature used hence number of data points.

**RESULTS**

**Exploratory Analysis**

There have been several studies in past which compare call log data (behavioral data) and survey data (cognitive data) and argue which explains the reality best (e.g. [15, 22]). These studies have reported general consistencies but specific misalignments in values such as the frequency of mobile phone use [3, 4, 15]. As a first step in validation, we started by comparing log data with the self-report measures using pairwise Pearson’s correlations. (Refer Table 2.)

The results show that the self-report measures are significantly correlated with the log data (p-values are less than 0.05 for 12 of the 15 comparisons). We can see that correlation r-values vary between 0.07 and 0.69 (median=0.3826). While this level of correlation would have been considered practically useful if we were considering two different phenomena, the associations are quite weak for variables which essentially measure exactly the same behavior. Hence we posit that these results suggest that the two data types are similar, but not a replacement of each other.

Unlike previous literature though (e.g. [3, 30]), we suggest that both types of variables have their own significance and instead of considering replacements, they should be used simultaneously. While they quantify similar things, they are different in their manifestations, and one cannot be neglected at the cost of other.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Last Day</th>
<th>Last Week</th>
<th>Last Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Calls</td>
<td>0.6233 (0.0001)</td>
<td>0.6900 (0.0001)</td>
<td>0.3290 (0.0142)</td>
</tr>
<tr>
<td>Number of people contacted</td>
<td>0.1247 (0.3644)</td>
<td>0.3137 (0.0197)</td>
<td>0.1927 (0.1587)</td>
</tr>
<tr>
<td>Number of calls to 1st frequently contacted person</td>
<td>0.3826 (0.0039)</td>
<td>0.0741 (0.5909)</td>
<td>0.1711 (0.2117)</td>
</tr>
<tr>
<td>Number of calls to 2nd frequently contacted person</td>
<td>0.6991 (0.0001)</td>
<td>0.4323 (0.0010)</td>
<td>0.4009 (0.0024)</td>
</tr>
<tr>
<td>Number of calls to 3rd frequently contacted person</td>
<td>0.4793 (0.0002)</td>
<td>0.4569 (0.0005)</td>
<td>0.3610 (0.0068)</td>
</tr>
</tbody>
</table>

Table 2. Correlations between call-log & survey-reported data. (P-values reported in brackets.)

**Predicting Social Outcome Variables**

An important reason for collection of self-reported data in longitudinal studies is to identify its associations with social outcome variables of interest. For example, [31] studies the interconnections between phone use, stress, and depression. Predicting the right outcomes for different individuals clearly be quite important. For example a better prediction of stress or depression could allow for the right treatment or intervention to help the affected individual.

In this study, we focus on three outcomes that can be of interest in a wide variety of social studies - social capital, social support and trust [11, 24, 29] and study the ability of different versions of the social data (self-reported vs logged) at predicting these outcomes of interest. Specifically, we ran regressions using self-reported variables, call-log variables, their differences and the combined dataset. We used the LASSO regression technique as we had a relatively modest number of data points. Unlike standard linear regression, LASSO adds penalty for each additional feature used hence we cannot attribute the improvements in predictive powers directly to the addition of more features [26].

The correlation scores (between predicted and actual outcome variables) and the Root Mean Square Error (RMSE)
are presented for four different types of regressions in Table 3. We ran different regressions with different sets of variables to test the importance of each type of data for prediction. For example, a model based on call-log variables had all the features collected through call logs and survey reported variables had similar features collected through surveys (calls per day, number of people contacted etc.). Difference variables were calculated by taking the absolute value of the difference between call-log data and survey reported data of the variable divided by the maximum of both. This is the operationalization of the aforementioned dissonance coefficient, one for each measured attribute.

\[
\text{Dissonance Coefficient} = \frac{|\text{Actual} - \text{Reported}|}{\text{Max(Actual,Reported)}}
\]

Lastly, we pooled all the variables together to predict social variables.

| Call-log variables | 0.4248 | 65.57 |
| Survey Reported variables | 0.5844 | 55.27 |
| Difference variables | 0.6633 | 44.45 |
| All variables combined | 0.9027 | 15.69 |

**Table 3(a). Regression Models to predict Social Capital**

| Call-log variables | 0.2633 | 196.35 |
| Survey Reported variables | 0.4953 | 159.43 |
| Difference variables | 0.6094 | 132.25 |
| All variables combined | 0.8569 | 62.65 |

**Table 3(b). Regression Models to predict Social Support**

| Call-log variables | 0.5153 | 107.81 |
| Survey Reported variables | 0.4373 | 118.74 |
| Difference variables | 0.4968 | 110.50 |
| All variables combined | 0.7855 | 57.79 |

**Table 3(c). Regression Models to predict Trust**

As we can see in Table 3 (a) and (b) that survey reported variables show more correlation and less RMSE than call-log variables, indicates that survey reported variables can have strong predictive power in predicting social variables and cannot be neglected or replaced by call-log variables.

From the above regressions, we can further infer that the difference between these two types of variables (behavioral and cognitive) carries significance in predicting the variables of interest. These “dissonance coefficients” capture the difference between what we think or perceive and what we actually do in reality.

When all these variables are combined, those models have highest correlations and predictive power for all three social variables of interest. Note that the increase in performance of the combined model cannot simply be attributed to more features as the LASSO approach selected a comparable number of features for each model. Hence, we interpret the results to mean that the models having all three sets of variables explain social capital, social support and trust better than any other model where these set of variables were taken individually.

**DISCUSSION**

The current study also has some limitations. First is the homogeneity of the sample. While this limitation prevents us from generalizing the findings to larger populations, the homogeneity allowed us to isolate social behavior as a predictor. A second limitation is the relatively small sample size - 47 participants. Taking into account these two limitations, we will be cautious in generalizing the results to larger populations until they are verified at scale.

Despite the limitations, this paper makes an important contribution towards the literature surrounding methodologies for understanding human behavior. It adds to the literature studying overlaps between self-reported and logged data and suggests two new viewpoints.

First, it argues a case for the two types of data (self-reported and logged) to be considered as complementary rather than replacements of one other. Both these types of data have been used in thousands of studies in fields ranging from sociology to interaction studies and human computer interaction. As evidenced here, a shift in the viewpoint on how to capture and interpret the data could yield very different insights into human behavior as well as their predictive ability over outcomes of interest. Even a small shift in perspective could imply significant changes in the findings obtained over decades of literature. For example, studies reporting (no) interconnections between call logs and social support may need to be revisited and the predictive effect sizes attributed to phone based metrics may change drastically. (For example, change from 0.26 to 0.86 as shown in Table 3(b)).

Second, it defines a newer metric – dissonance coefficient – and empirically demonstrates the important predictive role it plays in inferring multiple outcome variables of interest to social computing researchers. While there is prior literature connecting “optimism bias” and “positive interpretation bias” with psychological wellbeing [14, 23], our work suggests a different connection. Rather than identifying the positive or negative associations, our work interprets results from a methodological viewpoint and suggests that the consistency coefficient may yield significant predictive power on the social outcome variables of interest.

**CONCLUSIONS**

Although call-log variables are exact, when predicting the social variables like trust, support or social capital, which are nothing but a perception of an individual about himself/herself in society, we should not neglect the self-reported variables as they reflect the perception of one’s communication. In future studies, the self-reported or survey variables should not be looked as the replacement for log variables which undoubtedly has more accurateness but can act as supplement to predict these social factors more accurately.
REFERENCES


5. Brewer, D. D. et al., 1998. Comparison of direct estimate and partner elicitation methods for measuring the number of sexual and injection partners.. In 126th Annual Meeting of the American Public Health Association..


