Propitious Aggregation: Reducing Participant Burden in Ego-centric Network Data Collection

Derek Lackaff
Department of Radio-Television-Film
University of Texas at Austin
lackaff@mail.utexas.edu

ABSTRACT
One of the central challenges of ego-centric or personal social network research is minimizing the quantity of data that is requested from research participants while ensuring high data accuracy and validity. In general, collecting data about increasingly larger ego-centric networks places an increasing burden on respondents. The web-based Propitious Aggregation of Social Networks (PASN, http://pro.pitio.us) survey instrument reduces this burden by leveraging network data already available in the context of social network websites, and by providing an intuitive click-and-drag interface for survey responses. An experiment was conducted (N = 85), and the PASN method was found to produce networks which were significantly larger and more diverse than those produced using standard survey methods, yet required significantly lower time investments from participants.

Author Keywords
Social networks, social network sites, computer-assisted self interviews, methods, user interfaces.

ACM Classification Keywords
H.5.2 Information interfaces and presentation: Theory and methods; J.4 Social and behavioral sciences: Sociology.

General Terms
Design, Experimentation

INTRODUCTION
Researchers who study ego-centric (or personal) networks attempt to understand how the structure of an individual's social relationships results in meaningful outcomes. Personal networks are comprised of an *ego,* or focal individual, *alters,* or social contacts of ego, and the relationships that exist among these actors, which are termed *edges*. In general, collecting data about increasingly larger ego-centric networks places an increasing burden on respondents. Simply remembering the people in one's network can be surprisingly difficult [3], and providing detailed information about these alters and their relationships can be an overwhelming task to request of respondents [6]. Various survey-based "network generators" have been used in previous research to enable respondents to at least estimate the characteristics of their overall personal networks. This paper presents a method of personal network generation and interpretation that leverages previously-articulated social network data.

While research respondents are generally able to recall their more intimate relationships with a reasonable level of accuracy, less intimate relationships are more likely to be forgotten. Findings about informant accuracy have been mixed, and many researchers believe that informant accuracy is a topic of methodological interest [2, 4]. One study found that respondents were unable to recall 3% of best friends, 9% of close friends, and 20% of other friends [3], while another study found that 26% of close friends were forgotten [1]. Participant forgetting also has effects on structural measures of the networks [3]. Researchers generally use instruments called *name generators* to aid respondents in remembering and describing their networks. Name generators consist of a series of questions that elicit named alters as responses. Traditional name generators first use a general question, e.g. "Who are your closest friends?" then ask for relevant details (relationship, reasons for closeness, communication frequency, etc.) about each name that is provided [5].

Participant burden is a central concern in the design of personal network studies [10]. While a survey or interview may elicit a short list of important alters with just a few minutes of effort, larger networks can involve significantly more time. Data collection can also be repetitive and monotonous for the respondent – after remembering a number of family, friends, and acquaintances, the respondent is generally asked to interpret each of these names in some way relevant to the researcher's project. But perhaps the most intractable burden that is placed on participants is the evaluation of alter-alter relationships in their network: Does alter A know alter B? Does A know C? and so on. The geometric increase of potential alter-alter ties to be evaluated as networks increase in size can quickly become a crushing burden upon participants. A network of

15 alters contains 105 potential relationships, while a network of 50 alters contains 1,125 potential relationships.

Web-based computer-assisted self-interview (CASI) applications present substantial opportunities for personal network data collection. Web-based instruments can efficiently reach broad populations and generate large sample sizes, and can reduce researcher burdens of expense and data entry [7]. Web-based surveys can be designed to be highly visual and interactive, qualities that enhance respondent interest and data quality [8]. While personal network researchers must carefully consider a range of issues when implementing web-based instruments including visual layout, response burden, and sampling issues [14], the web presents a promising platform for the collection of high-quality network data [9]. Previous work with web CASI interfaces found that a majority of college-age respondents preferred a visual, dynamic interface to a standard text-based interface [11].

SOCIAL NETWORK SITES

Many web users are already quite familiar with exploring, visualizing and considering the structure of their personal social networks. Social network sites (SNSs) typically allow an individual to connect their personal profile to the profiles of other users, resulting in the public articulation of one's entire (online) social network. While the overlap of a SNS network and the corresponding "real" social network may be uncertain, the ready availability of SNS network data may present useful research opportunities. Recent web trends related to data transparency and interoperability suggest that the benefits of social network site data – e.g. the articulated network and the familiarity of respondents with the SNS interface – can be combined with CASI techniques to produce rich, meaningful personal network datasets. Vehovar et al.'s study of the impacts of various interface properties [14] concluded with the following design suggestions: simple, familiar web forms that help users understand question meanings as quickly as possible, paired with non-leading questions; using a single name box in name generation, and providing a dynamic interface that confirms the addition of each name, as well as provides the option to continue; to carefully consider the use of graphic elements, which may have powerful and unintended consequences for data collection; and to pretest data collection interfaces extensively. This paper presents an implementation of a network survey web application that builds on these findings.

SNS Application Programming Interfaces

In the summer of 2007, Facebook opened an application programming interface (API) to outside software developers (http://developer.facebook.com). Once authorized by a user, an application can access elements of that user's profile, such as personal information and network structure. The Propitious Aggregation of Social Networks (PASN) survey instrument is a web application that allows respondents to interactively generate and qualify personal network data. The instrument aggregates social network data that participants have already – propitiously – provided in the context of their use of social network sites. Importantly, PASN allows this data to be annotated and supplemented by research participants. PASN is available for download from http://pro.pitio.us. While PASN currently only integrates with Facebook, is will be straightforward to extend this approach to other SNSs such as Twitter, LinkedIn, or MySpace.

Integrating the data collection instrument with an extant social network data source like Facebook may provide several advantages: improved recall of network members [2], improved accuracy of network data, faster participant processing of large networks, and increased participant engagement in the research process. PASN displays alter thumbnail images (generally faces) in alter selection tasks, which appears to be a novel method in personal network data collection. An increasing body of evidence suggests that human faces are processed by the brain differently from other types of visual stimuli [13]. While the implications of this have not been explored empirically (although [8] provides anecdotal evidence to this effect), it seems likely that the visual presentation of alters may allow respondents to process alter lists more quickly and naturally. The key issues to be explored relate to expected differences in the networks produced by research participants using different kinds of instruments. Having a pre-generated list of relationships should facilitate memory of more specific alters. Therefore,

H1. Participants using the social API instrument will identify more key discussion partners (alters) than those using standard name-generation techniques.

Second, people who have larger network of contacts on a social networking site are likely to have a larger proportion of their key discussion partners represented as contacts. Therefore,

H2. Participants with larger articulated social networks will need to manually enter fewer alters than those with smaller articulated social networks.

The API instrument should allow participants to more easily recall alters from multiple social contexts. *Density* is a property that refers to the proportion of edges present in the network. In this case, alters will be less likely to know one another. Therefore,

H3. Networks generated using the social API instrument will produce networks that are less dense than those generated manually.

As previously noted, a primary concern of personal network name generation is the duration of name generation endeavors. Therefore the research question,

RQ1: How much time does the drag-and-drop edge generation method require from participants?

Methods

This study was designed to compare the best practice single-box name generation method identified by [14] with the PASN method. Participants were recruited from several sections of a large, introductory Communication class in accordance with IRB regulations, and received class credit for their participation. A link to the survey was posted by the instructor on a class website, and participants enrolled in the study by clicking this link. Participants were then assigned to one of two conditions.

The Standard Condition name generator consisted of six questions about "key discussion partners" from various life domains (e.g. family, friends from high school, work colleagues). Each question was followed by a single text-entry box, into which the participant typed a name. Participants were encouraged to enter as many alter names as they could remember. After all six name generation items were completed, the participant began the edge creation task, a procedure to determine how the participant's identified alters know one another. Edges were defined by presenting the participant with a question: "Which of these people does <alter name> know?" The participant then clicked and dragged alters icons from the name generation task to an answer area at the top of the interface. Several strategies were employed to reduce the time and effort required of participants. First, all relationships were assumed to be reciprocal, immediately halving the number of potential edge questions. Further, once the edge list had been defined for a given alter, that alter icon was removed from the interface in future questions.

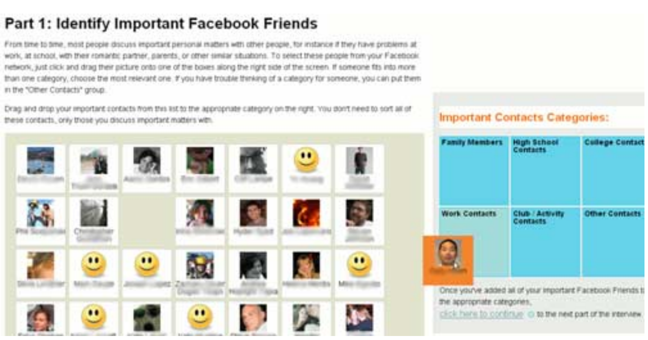

Figure 1. PASN Facebook API name generation interface.

Participants in the API Condition were presented with a slightly different procedure. First, they were prompted by the application website for their Facebook username and password. Next, they were presented with the interface shown in Figure 1. Their Facebook friends were presented as "icons" containing the friend's name and Facebook profile picture, and the participant was prompted to drag the icons of key discussion partners to the appropriate category on the right.

Following this identification of important Facebook friends, the participant answered the six name generation questions as in the Standard Condition discussed above. The edge generation task was then completed as in the Standard Condition, with all alters (those entered manually and those imported from Facebook) presented together in their relevant groups.

Results

A total of 85 individuals participated in the study. 48 participants were female, and their mean age was 23 (SD = 4.30). Six participants maintained profiles on SNSs other than Facebook. With the exception of these 6 participants, who were assigned to the Standard Condition, participants were randomly assigned to the two conditions, resulting in 38 total participants in the Standard Condition and 47 in the API condition.

The first hypothesis predicted that larger networks of important alters would be identified by users of the API-integrated instrument. To test this hypothesis, t-statistics were calculated for each of the API groups. On average, participants articulated larger networks using the API interface (M = 59.24, SD = 13.87) than participants using the standard interface (M = 8.62, SD = .80). This difference was significant $t(46.28) = -3.67$, $p < .01$ (unequal variances). Thus the first hypothesis was supported. Since the task was identical in each condition – to identify the people with whom one discussed "important matters" -- this difference indicates a substantial impact of the network generation interfaces. The mean network size of 8.6 important alters is in line with traditional size findings of "strong tie" networks, and provides a replication of [14]'s findings. However, the much larger networks generated using the API data (a mean of nearly 60 alters) indicate that the definition of "key discussion partner" may be highly contextual with regards to the data collection instrument employed. When presented with a list of friends' names and photographs, many more salient discussion partners can be identified.

The second hypothesis predicted that larger SNS networks represent broader coverage of the overall social network, suggesting that participants with larger SNS networks would need to manually identify fewer important alters. Both the Facebook friends variable and the manual friends variable were significantly skewed, so these variables were log normalized for further analysis. When an ordinary least squares linear regression of the number of Facebook friends on the total of manually-added alters was performed, the model was significant ($\beta = -.38$, $F(1,45) = 6.92$, $p < .05$). This suggests that the more Facebook friends one has, the fewer alters that must be manually entered to complete the "important people" name generation task. This supports the second hypothesis. If a participant has a large Facebook network, then there is a proportionally smaller quantity of key discussion partners who are not Facebook friends. This bodes well for studies that utilize API-integrated instruments within populations of heavy social media users.

The third hypothesis predicted that the networks generated in the Social API condition would be less dense than those produced in the standard condition. Mean densities were indeed lower in the API condition (M = 0.21, SD = 0.13) than in the standard condition (M = 0.06, SD = 0.05). This difference was significant, $t(44.23) = 6.39$, $p < .001$ (unequal variances). Further, a regression model controlling for network size also supports this conclusion, $F(2,82) = 26.82$, $p < .001$, with the API condition (final $\beta = -.55$, $p < .001$) explaining 26.5% of the variance in the model. Traditional name generation often results in relatively dense networks, as alters that are freely-recalled are often produced in groups (family members, coworkers, activity partners, etc.). As the relationships maintained on social network sites often cross multiple social domains and contexts, less biased networks may be produced using API methods.

The research question was concerned with the amount of time participants spent generating the edges of the personal network. Large differences between the two conditions were apparent. Participants in the standard condition identified an average of 8.5 important alters (SD = 5.00) and completed the name generation task in 2.5 minutes (SD = 1.80). Participants in the API condition identified an average of 62.4 important alters (SD = 100.40) and completed the name generation task in 8.7 minutes (SD = 4.73). In the API condition, the maximum time spent was 21 minutes, by a participant who identified 653 ties among 171 alters. However, one participant identified 1,265 edges among 472 alters in 19 minutes. This extreme variability does not allow for a conclusive statement to be made, but suggests that the PASN method may allow participants to articulate and define large personal networks in reasonably short time periods.

The edge generation interface used by the web application presented alters for evaluation at the network level, rather than the alter level. This meant that all possible ties for a given alter were determined in a single task, rather than requiring the evaluation of potential relationships as separate tasks, as in the interface used by [12]. This resulted in a much lower completion time for the edge generation task. If N equals the number of alters in the network, [12]'s strategy would have presented the participant with $N * (N - 1) / 2$ queries. The PASN interface presented the participant with N queries. This allowed networks containing hundreds of alters to be processed in minutes, rather than hours. It is important to note that edges were defined here in a very general manner (who knows whom?) and nothing is known about the quality of these relationships (e.g. good friends, casual acquaintances, bitter enemies etc.) Edges may of course be defined more specifically, at the cost of requiring the participant to make more thoughtful and informed evaluations. There are concerns about false negatives (relationships that exist, but that were not identified), and future work will evaluate their prevalence. Overall, the benefits of this approach in terms of reduced participant burden appear to greatly outweigh the threats to validity of the collected data.

ACKNOWLEDGMENTS

The author gratefully acknowledges the technical contributions of Jeffery Fritz and Joseph Hsu and the helpful suggestions of four anonymous reviewers.

REFERENCES

1. Bell, D. C., Belli-McQueen, B., & Haider, A. Partner naming and forgetting: Recall of network members. *Social Networks 29* (2007), 279-299.
2. Bernard, H. R., Killworth, P., Kronenfeld, D., & Sailer, L. The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology 13*, (1984), 495-517.
3. Brewer, D. D., & Webster, C. M. Forgetting of friends and its effects on measuring friendship networks. *Social Networks 21*, (1999), 361–373.
4. Butts, C. T. Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks 25,* (2003), 103-140.
5. Fischer, C. *To dwell among friends.* Chicago: University of Chicago Press. 1982.
6. Fu, Y.-C. Measuring personal networks with daily contacts: A single-item survey question and the contact diary. *Social Networks 27, (*2005), 169-186.
7. Gosling, S. D., Vazire, S., Srivastava, S. John, O. P. Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist 59,* (2004), 93-104.
8. Hogan, B., Carrasco, J.-A., & Wellman, B. Visualizing personal networks: Working with participant-aided sociograms. *Field Methods 19,* (2007), 116-114.
9. Lozar Manfreda, K., Vehovar, V. & Hlebec, V. Collecting ego-centered network data via the web. *Metodološki zvezki 2*, (2004), 295-321.
10. Marin, A. & Hampton, K. N. Simplifying the personal network name generator: Alternatives to traditional multiple and single name generators. *Field Methods 19*, (2007), 163-193.
11. McCarty, C., & Govindaramanujam, S. A modified elicitation of personal networks using dynamic visualization. *Connections 26,* (2005)*,* 61–69.
12. McCarty, C., Killworth, P. D., & Rennell, J. Impact of methods for reducing respondent burden on personal network structural measures. *Social Networks 29*, (2007), 300-315.
13. Palermo, R., & Rhodes, G. Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia 45*, (2007), 75-92.
14. Vehovar, V., Lozar Manfreda, K., Koren, G., & Hlebec, V. Measuring ego-centered social networks on the web: Questionnaire design issues. *Social Networks 30*, (2008), 213-222.