Inferring Network Structure Relationship over Mobile Phone Data

By

Choompol Thongchuenjitt

To

Prof. Eduard Heindl

BCM Summer Semester 2012

21st June 2012
All intellectual information mentioned in this paper has been cited and paraphrased. It is to be declared that all the work done in this document is solely and explicitly by the author of this document.

21st June 2012

Mr. Choompol Thongchuenjitt
# Table of Contents

1. Introduction .................................................................................................................. 4
2. Observational Data........................................................................................................ 4
   2.1 Advantages ............................................................................................................. 4
   2.2 Disadvantages ....................................................................................................... 5
3. Self-report Data ........................................................................................................... 5
   3.1 Advantages ............................................................................................................. 5
   3.2 Disadvantages ....................................................................................................... 5
4. Examine relationship based on observational data ....................................................... 5
   4.1 Define relationship on recency ............................................................................. 6
   4.2 Analyze whether behaviors data identified in observational data or not ............... 7
   4.3 Study the relationship between behavior data and individual satisfaction .......... 11
5. Conclusion .................................................................................................................... 13
Table of Figure

1. Figure1 - The effect of recency on self-report data ................................................................. 7
2. Figure2 - Probability of Proximity .............................................................................................. 8
3. Figure3 - Factor Analysis Loadings ............................................................................................ 9
4. Figure4 - Normalized extra-role histogram .............................................................................. 10
5. Figure5 - The Extra-role Factor ............................................................................................... 11
6. Figure6 - Friendship Network .................................................................................................. 12
1. Introduction
Nowadays, there are many mediums for people to use to communicate to each other such as mobile phone, emails, Short Message Service, instant messaging, and social network. In the past, people use to communicate to each other via sending letters, sending postcards, landlines calling, and conversation. Comparing to the past, it is easier to contact other people globally in present.

The interaction of face-to-face among people seems to decrease rapidly due to the improvement of technology that helps people to communicate to each other despite the locations and time. Therefore, there are many studies of interactions among people via remote communication to find the importance of itself whether this can be replaced the face-to-face interaction or not.

The comparison of inferring friendship between observational data from mobile phone and self-report survey data is difficult because it is hard to compare these two types of data. For example, it is easy to distinguish the relationship based on self-report survey data than mobile phone data depending on the interaction whether it is new and unique or not. Also, it is possible to accurately infer 95% of the relationship based on the observational data while they are performing distinctive temporary and spatial patterns of physical proximity and calling patterns. Therefore, these behavioral patterns can be used to determine the individual level of outcomes such as job satisfaction1.

2. Observational Data
The rapidly growing volume of electronic communication data, such as email exchange, instant messaging, mobile phones, online games, social networking or social media sites, has been a great benefit to social network analysis, enabling researchers to study networks at very large scales and over extended time periods. However, the excitement generated by this explosion of available data has overshadowed two distinct but related problems:

1. The inference problem, those “real” social ties are not directly observable and hence must be inferred from observations of events, like physical interactions or communication records.
2. The relevance problem, that there is no one “true” social network, but rather many such networks, each corresponding to a different definition of a tie, and each relevant to different social processes2.

2.1 Advantages
Validity – With this method in collecting data, there will be less bias than interviewing. Moreover, people will act naturally when they feel unnoticeable.

---

1Inferring Friendship Network Structure, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/#B1
2 Inferring Relevant Social Networks from Interpersonal Communication, http://research.yahoo.com/pub/3168
Quantitativeness – With this method, the data can be collect as much as possible due to the constrains which are time and efforts

2.2 Disadvantages

Accuracy – Human’s feeling is unpredictable and complex. Sometimes, it depends on the surrounding

Security – Most people feel uncomfortable to be observed without permission by someone. Therefore, the security could be an important issue to collect observational data.

3. Self-report Data

A self report is a method which involves asking individual about the attitude, ideas, beliefs and feelings. Moreover, self-report are often used as a way to gain responses in studies and experiments or, in another word, in science or business perspectives. There are many methods for collecting this type of data which are questionnaires, interviews, yes-or-no questions, open-and-close questions, and rating scales.

3.1 Advantages

Quickness and ease – it is easy for people to answer the yes-or-no question, and people tend to come up with answer when there is a face-to-face interactions.

Detailing – With face-to-face interactions, the interviewers can easily read interviewee’s gesture before making continuous questions for more precise information.

3.2 Disadvantage

Complexity – Sometimes, people seems uncomfortable to do questionnaires or interviewing because of the complexity and difficulties of questions which could affect the validity of the information.

Social-desirability bias - The participants may answer the questions not truthfully to portray themselves as socially acceptable.3

4. Examine relationship based on observational data

Human interactions or social network analysis has been constrained in accuracy, breadth, and depth because of its reliance on self-report data. Social network studies relying on self-report relational data typically involve both limited numbers of people and a limited number of time points (usually one). As a result, social network analysis has generally been limited to examining small, well-bounded populations, involving a small number of snapshots of interaction patterns. Although important work has been done over the last 30 years to analyze

the relationship between self-reported and observed behavior, much of the social network literature is written as if self-report data are behavioral data.\textsuperscript{4}

The Reality Mining study followed 94 subjects using mobile phones preinstalled with several pieces of software that recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status. Subjects were observed using these measurements over the course of nine months and included students and faculty from two programs within a major research institution. We also collected self-report relational data from each individual, where subjects were asked about their proximity to, and friendship with, others. Subjects were also asked about their satisfaction with their work group. We will hereafter refer to data collected purely from mobile phones as “behavioral” data as opposed to “self-report” data.\textsuperscript{5}

4.1 Define relationship on recency

The reliability of existing measures for relationships has been the subject of sharp debate over the last 30 years, starting with a series of landmark studies in which it was found that behavioral observations were surprisingly weakly related to reported interactions. There are multiple layers of cognitive filters that influence whether a subject reports a behavior. Existing research suggests that people are good at recalling long-term, but not short-term, social structures. We examine whether there are systematic biases in recall that have been observed in other areas with respect to human memory, specifically, whether there are recency and salience biases in recall of physical proximity. A recency bias is one where memories are biased toward recent events. A salience bias is one where memories are biased toward more vivid events. Here, we capture recency by the quantity of interactions in a fixed period preceding the survey, and salience by whether the individual in question is a friend or nonfriend.\textsuperscript{6}

\textsuperscript{6} Inferring Friendship Network Structure, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/#B1
This figure shows that this recency effect peaks when we use a seven-day window for recent interactions, suggesting that individuals recall of information about their interactions begins to degrade after approximately one week.

4.2 Analyze whether behaviors data identified in observational data or not

Observing intrinsically cognitive relationships, such as friendship or love, is a fundamentally different challenge than observing whether two people are near each other. It is clear, for example, that two individuals can be friends without any observable interactions between them for a given period. Context, however, especially spatial and temporal, is likely to be an important indicator of particular types of relationship, where spending a couple of hours in close proximity at a location away from work on a Saturday night is quite different from spending a couple of hours in close proximity at work on a Wednesday afternoon, for example. Here, we borrow from cognitive science the idea of scripts. Specifically, we examine whether proximity, location, and time cluster together in a predictable fashion and whether these behavioral patterns, in turn, predict friendship.

---

7 The effect of recency on self-report data, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F1/
Probability of proximity is calculated for each hour in the week and is generally much higher for friends than non friends. However, it is also apparent that asymmetric and symmetric friend dyads have different temporal and spatial patterns in proximity, with symmetric friends spending more time together off campus in the evenings.

Figure2. Probability of Proximity

9 Probability of proximity, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F2/
For relationship inference, based on a promax rotation, it is possible to divide the dyadic variables into the two factors above: in-role and extra-role communication. In-role communication consists of the behaviors typically associated with colleagues whereas extra-role communication corresponds to more personal behavior such as proximity on Saturday nights or at home.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Specific variance</th>
<th>Factor 1: Extra-role</th>
<th>Factor 2: In-role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work proximity, weekdays, 8 a.m.–8 p.m.</td>
<td>0.005</td>
<td>-0.119</td>
<td>1.07</td>
</tr>
<tr>
<td>Work proximity, weekdays, 8 p.m.–8 a.m.</td>
<td>0.568</td>
<td>0.555</td>
<td>0.144</td>
</tr>
<tr>
<td>Work proximity, weekends</td>
<td>0.642</td>
<td>0.501</td>
<td>0.137</td>
</tr>
<tr>
<td>Off-campus proximity, weekdays, 8 a.m.–8 p.m.</td>
<td>0.310</td>
<td>0.691</td>
<td>0.195</td>
</tr>
<tr>
<td>Off-campus proximity, weekdays, 8 a.m.–8 p.m.</td>
<td>0.240</td>
<td>0.946</td>
<td>-0.123</td>
</tr>
<tr>
<td>Off-campus proximity, weekends</td>
<td>0.291</td>
<td>0.914</td>
<td>-0.119</td>
</tr>
<tr>
<td>Phone communication</td>
<td>0.806</td>
<td>0.469</td>
<td>-0.047</td>
</tr>
</tbody>
</table>

Figure 3. Factor Analysis Loadings\textsuperscript{10}

\textsuperscript{10} Factor analyzing loadings, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/table/T1/
In this histograms, the distributions of a pair of colleagues extra-role communication factor scores segmented by relationship. Ninety-five percent (21/22) of the symmetric friendships have extra-role scores above 5, whereas ninety-six percent (901/935) of symmetric non friends have extra-role scores below 5. The 28 asymmetric friends have more behavioral variance, drawing from behaviors characteristic of both non friends and friends.

Figure 4. Normalized extra-role histogram\textsuperscript{11}

\textsuperscript{11} Normalized extra-role histogram, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F3/
A histogram of the extra-role distribution generated from behavioral data collected from September to January for two sets of dyads. The red bins represent the dyads that consistently confirmed they were not friends on both the January and May survey (n = 2153). The yellow bins represent the dyads that confirmed they were not friends on the January survey, but at least one individual named the other as a friend on the May survey (n = 32). Clearly these two sets of dyads come from distinct distributions; potential explanations for the yellow distribution could be survey error in January (i.e., the friendships existed, but were not reported in January), or that the dyads' behavior during the autumn was indicative of budding friendships that they only became aware of during the subsequent year.

4.3 Study the relationship between behavior data and individual satisfaction

However, this likely understates the accuracy of the behavior-based inference of friendship, because the self-report measure for friendship itself is probably not perfectly reliable. To compare the validity of these two measures of friendship, we examine the effectiveness of both measures in predicting social integration in work groups.

Inferred, weighted friendship network vs. reported, discrete friendship network. Frame A shows the inferred friendship network with edge weights corresponding to the factor scores for factor 2, extra-role communication. Frame B shows the reported friendship network. Node colors highlight the two groups of colleagues, first-year business school students (brown) and individuals working together in the same building (red).
Conclusion

Data collected from mobile phones have the potential to provide insight into the underlying relational dynamics of organizations, communities and, potentially, societies. These methods provide, for example, a new approach to studying collaboration and communication within organizations allowing the examination of the evolution of relationships over time. There is no technical reason why data like these cannot be collected from millions of people throughout the course of their lives.

These methods in collecting observational data from more than million participants are more efficient than collecting self-report data, but there are some drawbacks from this type of data. Firstly, humans’ feelings are sensitive and unpredictable. Therefore, the observational data may not accurate as much as self-report data. Secondly, the privacy is a security issue that can interrupt to collect observational data.
References

Factor analyzing loadings, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/table/T1/

Friendship network, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F5/


Inferring Relevant Social Networks from Interpersonal Communication, http://research.yahoo.com/pub/3168


Probability of proximity, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F2/

The effect of recency on self-report data, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F1/

The extra-role factor, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2741241/figure/F4/
