Social Network Analysis Using Perceptual Tomography

A Thesis Presented in Partial Fulfillment of the Requirements for the Master in Science in Digital Forensics and Cyber Security

John Jay College of Criminal Justice

City University of New York

Ahmet Tamer Oguz

Summer : August 2014

Social Network Analysis With Perceptual Tomography

Ahmet Tamer Oguz

This thesis has been presented to and accepted by the Office of Graduate Studies of the John Jay College of Criminal Justice of the City University of New York in partial fulfillment of the requirements for the Master in Science in Digital Forensics and Cyber Security.

Dr.Bilal Khan		
Thesis Advisor	Signature	Date
Dr.Richard Lovely		
Second Reader	Signature	Date
Dr. Anne Lopes		
Dean of Graduate Studies	Signature	Date

ACKNOWLEDGEMENTS

It is a real pleasure for me to have reached this moment after the long process of writing this thesis. Here I would like to express my sincere thanks without whom this thesis would never been completed.

I would like to thank Turkish Gendarmerie General Command for providing me the financial support for my graduate studies. I present my gratitude to my commanders for encouraging gendarmerie officers to couple their expertise in the field with the academic education.

I owe very much to my thesis advisor, Dr. Bilal Khan, for his guidance, encouragement, and patience. Working with him has cherished my interest and enthusiasm on the research. I am really indebted to Dr. Khan for the long hours he devoted for my academic and personal development.

I would like to express my sincere thanks to Dr. Richard Lovely for his efforts for making the Digital Forensics and Cybersecurity Program a challenging, but a rewarding one. Along with his position as the program director, he was also a mentor for me to realize my ultimate goals in the program.

My deepest thanks go to my wife, Işık. During this difficult process, her love and patience have been the greatest motivation for me.

ABSTRACT

A novel data collection approach where a researcher simultaneously recruits respondents for sampling and collects relational data leveraging proximity perception is presented. The key idea underlying our approach is that by collecting social proximity information from respondents without requesting an enumeration of ego ties, we can achieve the collection of the relational data more efficiently without raising privacy and anonymity concerns. The main assumption of our approach is that whenever the geodesic distance between alters within a social network is not too large, respondents can perceive and report the distance value.

Starting from this assumption, we develop three models, and use simulations to evaluate their performance. For each of these models, we consider different sampling methods in particular, random sampling and Respondent Driven Sampling (RDS). In addition, we develop algorithms to organize the entities within the sample to efficiently elicit the perceptions of the respondents.

Our results indicate that this new approach is able to generate network distance estimates that are coherent with the underlying social network topology. With regards to sampling, we find that the new approaches presented here performs best when coupled with the RDS method.

Contents

List of Figures

List of Tables

1 Introduction

Wasserman and Faust (1994) define social network analysis as a research perspective which encompasses models, theories and applications that are expressed in relations of social actors in a social network. In the field of social network analysis, studies on hidden populations attract a growing interest because of their implications for public health and safety.

Hidden networks are communities whose activities are concealed from others because of social and legal sanctions (Hendricks & Blanken, 1992; Watters & Biernacki, 1989). Loosely linked cyber networks of hackers, child porn users, exotic usenet groups, and criminals who use online means are hidden networks which evolve in cyber space. Understanding the structure of these networks and locating important actors are vital to explain the relations within these networks or to conduct efficient investigation of cybercrime. However, classical approaches which study known social networks are insufficient when analyzing these networks (Lu, Polgar, Luo, & Cao, 2010; Holt, Strumsky, Smirnova, & Kilger, 2012).

One challenge in studying hidden populations is the unavailability of network data. Archival data, such as police, judicial and institutional records are difficult to access (Arsovska, 2012), and they are also prone to bias (Watters & Biernacki, 1989). In addition, traditional data collection methods such as household surveys, telephone and email surveys are usually inefficient because these hidden communities form a small fraction of the entire population (Heckathorn, 1997). Therefore, the researcher needs to access individuals systematically using more sophisticated sampling techniques (Muhib et al., 2001; Heckathorn, 1997).

Another challenging issue in research on hidden networks is how best to collect relational information. Because of the stigmatized nature of participation within these networks, privacy and anonymity are of particular concern, and these concerns can affect the quality and quantity of the respondents' reports (Arsovska,

2012). Researchers who seek to collect a maximum amount of accurate data using minimal resources should bear in mind these trade-offs, in order to minimize falsified or biased reports and to maximize participation by the target population (Spreen, 1992; Heckathorn, 1997; McNeeley, 2012; Dombrowski, 2012).

In this thesis, we develop an interview-based data collection design in which we sample individuals and collect network data simultaneously. Our method collects the *perceptions* of social proximity from studied respondents. Through simulations, we evaluate the performance of our data collection design, and interpret the results of these simulation experiments.

We demonstrate that this method is able to collect and synthesize accurate network data while raising fewer privacy and anonymity concerns. The method also produces estimates of relative social network distance between pairs of actors.

The document is organized as follows. In Section 2, we introduce the background on Social Network Analysis (SNA). We start by presenting an overview of fundamental concepts, then we describe several commonly used sampling methods, and present techniques for collecting relational network data, and both their strengths and weaknesses. In Section 3, we introduce the problem statement, our objectives, the main idea underlying our approach, and assumptions. Section 4 explains our methodology for evaluating the performance of the proposed design. Section 5 briefly presents our simulation environment. Section 6 presents the experiments we conducted using three different models. Section 7 discusses the results of these experiments, comparing the models under consideration. In Section 8, we present the limitations of our study, and describe possible future extensions to this work.

2 Background

The field of SNA analyzes social ties and network structures towards advancing the understanding of human behavior. The main proposition of SNA is that social actors and their networks are interdependent (McGloin & Kirk, 2010), and that social networks have implications on individuals actions because they provide both opportunities and constraints (Wasserman & Faust, 1994b).

SNA is an interdisciplinary field which includes both theoretical concepts and methodological techniques (Papachristos, 2011). SNA expresses theories, models and applications concerning patterns and regularities of relations between actors (Wasserman & Faust, 1994b). Its methodological techniques are drawn from many fields including graph theory, statistics, algebraic models (McGloin & Kirk, 2010; Papachristos, 2011), simulation and visualization (Marsden, 2005).

2.1 Social Networks

The main components of a social network are actors and links. Actors can be any social units, such as people, institutions or web sites. Links are the pairwise social ties or relations channels among pairs actors, e.g. contacts, group attachments or meetings (Scott, 2000).

Graph theory provides a formal language and a set of techniques to study social networks. In network terms, a social network is denoted as $G = (V, E)$, where V is the set of vertices or nodes (actors), and E is the set of edges, ties or links (relations). The number of elements in V is the size of the network, which usually denoted as N.

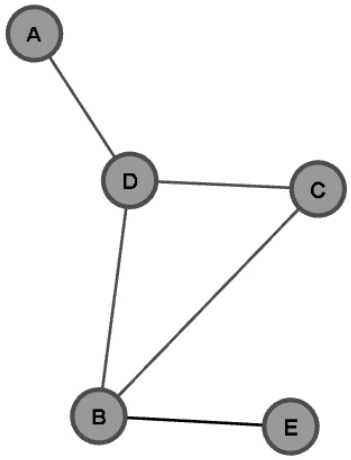


Figure 1. A graph with 5 nodes

For example, in Figure 1, there is a five-node graph where $V = \{A, B, C, D, E\}$ and $E = \{(A, D), (B, C), (B, D), (B, E), (C, D)\}$.

The main objects of study in SNA are actors and their relations, such as dyads and triads (Wasserman & Faust, 1994b). A Dyad is a subgraph which consists of a pair of vertices and a tie between them. A triad is formed by three nodes and the existing ties connecting each other.

According to the types of relations, there are two types of graphs: directed and undirected (Wasserman & Faust, 1994a). If the relation is oriented from one actor to another, there is a sender-receiver relationship between actors, and then the resulting graphs are referred to as directed. In undirected graphs, there is no such orientation implied.

The **geodesic distance** is the shortest path distance between a pair of vertices (Wasserman & Faust, 1994a). In total, $N \times (N - 1)$ ordered vertex pairs in an N-node graph, and there are $\binom{N}{2}$ unordered vertex pairs.

Let u and v be two distinct vertices $\in V$, and $d(u, v)$ be the function that returns the length of shortest path distance connecting u to v. There can be two distinct scenarios.

Scenario I: d(u,v) is a a finite number. Note that if d(u,v) = 1 then u and

v are connected by an edge . If $d(u, v) > 1$, there is no edge between u and v , but there is a finite path connecting these vertices.

Scenario II: If d(u,v) is infinite. It implies that one can not reach from u to v.

Based on geodesics, one may define properties such as centrality and density (McCormick, 2011; Papachristos, 2011). If there are many pairs with short geodesic paths connecting them within the social network, this indicates that the network is "centralized". On the other hand, a non-centralized network has a few long chains (McCormick, 2011). Furthermore, when the geodesic distance between a large fraction of pairs of nodes is finite, this indicates strong connectivity within the network (Wasserman & Faust, 1994a).

The number of geodesic paths which involve an individual actor is the basis of the actors' **betweenness centrality**, and the sum of geodesic distances from other vertices to an individual node indicates the **closeness** of that node.

Network density measures the average strength of the connections (Marsden, 1990). It is taken to be the proportion between actual links in the network to the theoretical maximum which is $\binom{N}{2}$.

The number of ties of an individual node is called its **degree**, and is one of the most common measurement used to identify important actors in a social network. From the network perspective, degree distribution of actors is an important measure.

Having given an overview of SNA, we will now discuss sampling and data collection techniques. The reader who is interested in further details on SNA is referred to Wasserman and Faust, (1994), Scott, (2000), and Butts, (2008).

2.2 Sampling Methods

Broadly speaking, we can categorize sampling methods into two classes: probabilistic sampling and non-probabilistic sampling (Meyer & Wilson, 2009; Heckathorn, 1997).

2.2.1 Probabilistic Methods. In probabilistic methods, each actor in the social network is included in the sample with a non-zero probability (Sudman, 1976). There are several methods for random sampling, namely simple random sampling, stratified and cluster random sampling. In simple random sampling, every actor in the population has the same probability of being included in the sample, whereas in stratified and clustered sampling, different subgroups may be assigned different probabilities for being selected (Meyer & Wilson, 2009).

The greatest advantage to the probability sampling is that it allows the researcher to generalize findings from the sample to the entire population. However, when network being considered is large or the target is comprised of hidden populations, these techniques may not be viable. In such a circumstance, the number of subjects is sparse within the population as a whole, and because of this, the researcher must screen a large segments of the population, which is infeasible due to the costs involved (Watters & Biernacki, 1989; Stueve, O'Donnell, Duran, Doval, & Blome, 2001).

2.2.2 Non-Probability Sampling Methods. In non-probability sampling techniques, the probability of a subject being included in the sample is not known. Thus, in these techniques the sample may exclude certain some sub-populations or be otherwise biased.

In these techniques, the researcher specifies the aims and the purpose of the research, and characterizes the network of interest. Then the researcher designs a sampling method in a convenient and efficient way, with the purpose of locating subjects (Hendricks & Blanken, 1992). Researchers also refer this sort of sampling

techniques as convenience sampling (Meyer & Wilson, 2009).

Below we will explain some widely used non-probability methods: community venues, time-space sampling, snowball sampling, and finally, respondent-driven sampling.

2.2.2.1 Community Venues Technique. In this method, the researcher samples people from locations where the community of interest is expected to be present. Potential venues may include medical institutions, counseling centers, shelters, local bars etc (Miller, Wilder, Stillman, & Becker, 1997; Meyer & Wilson, 2009).

2.2.2.2 Time-Space Sampling. Time-space sampling is a three step process. First, a sample of venues is selected from the universe of venues, and attendance statistics are screened according to time intervals. Second, weights are assigned for specific time-space venues to select subjects. Finally, venues are visited at the specific times selected, and attendees are recruited to participate in the survey (Stueve et al., 2001; Muhib et al., 2001).

2.2.2.3 Snowball Sampling. Snowball sampling is a method which relies on social networks to locate subjects. After finding the initial participants with previously mentioned techniques, the researcher asks each of the initial respondents to report other people in the subject pool of interest. In this way, the researcher uses previous participants to discover additional participants in the study (Biernacki & Waldorf, 1981; Hendricks & Blanken, 1992; Wasserman & Faust, 1994a; Petersen, 2005).

2.2.2.4 Respondent Driven Sampling (RDS). RDS is a dual-incentive chain-referral sampling method which relies on social networks to find recruits (Heckathorn, 1997). This method assumes that members of a hidden population are better able to contact their peers for recruitment into the study than the researchers themselves.

The RDS method begins with recruitment of a number of respondents, which are referred to as "seeds". The primary incentives are given to the seeds when they participate in being interviewed. When a seed completes his/her interterview, a fixed number of "RDS coupons" are given to them, and the interviewee is offered a secondary incentive for each new subject who participates in the survey process by presenting a coupon. The dual-incentive process is applied recursively to new participants in the same manner as for the original seeds (Heckathorn, 1997, 2002; Salganik & Heckathorn, 2004; Ramirez-Valles, Heckathorn, Vazquez, Diaz, & Campbell, 2005).

Prior work indicates the RDS method can yield large sample sizes and provides a good fit to archival records of hidden communities (Salganik & Heckathorn, 2004; Dombrowski, Khan, Moses, Channell, & Dombrowski, 2012).

RDS is different from the snowball sampling, for the participants do not disclose their peers to the researcher directly. Instead new participants are recruited by previous participants, and participate in the survey willingly. In this manner, RDS raises fewer privacy and anonymity concerns. Moreover, research shows that respondents are more likely to participate in the study when they are recruited by previous participants with whom they have close ties (Wejnert, 2009).

2.3 Techniques to Collect Relational Data

In this part, we will introduce techniques used in surveys and interviews to collect relational data.

2.3.1 Whole Network Studies. In whole network surveys, the roster is the most commonly used. (Butts, 2008; Marsden, 2005). In this method all members in the sample are enumerated in a roster, and then respondents are asked to report their ties to other members. This is a convenient method for respondents because they do not need to remember or enumerate the members in their social

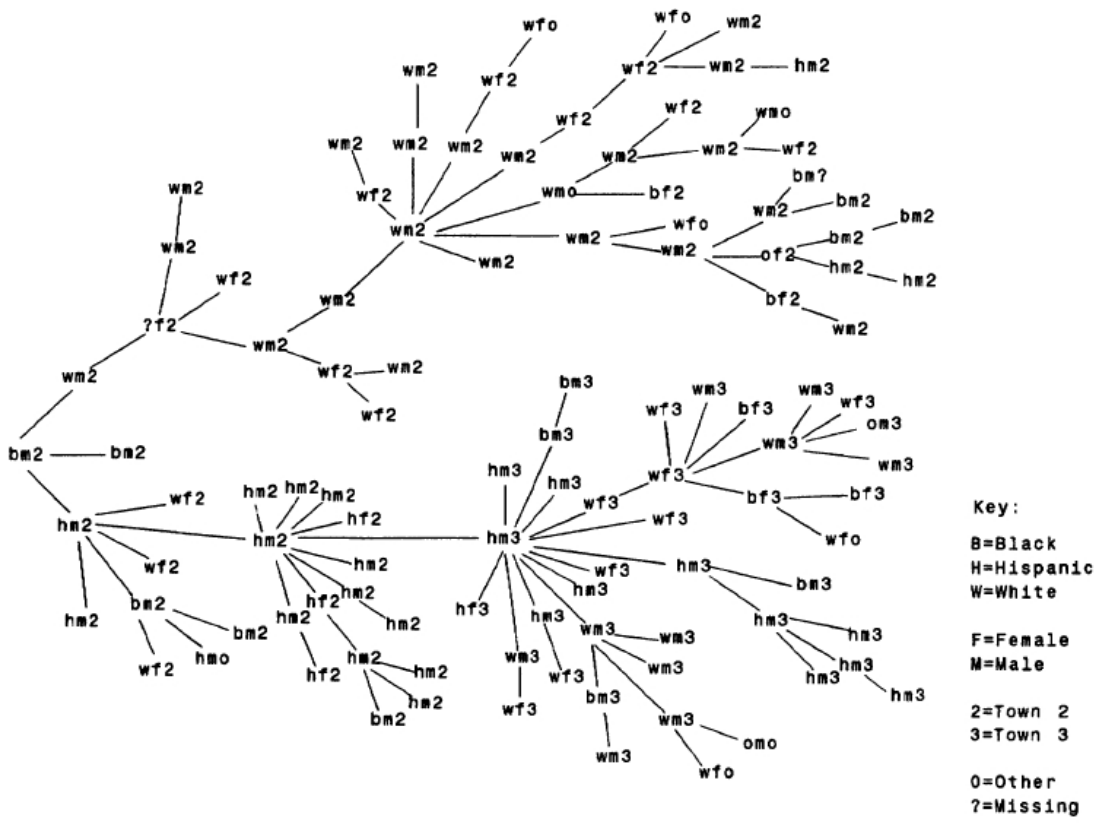


Figure 2. Recruitment network of Respondent Driven Sampling from Heckathorn, (1997).

network. However, this method requires complete knowledge of the members beforehand, and may also be impractical when the network of interest is too large and/or represents a hidden population. It also requires a follow-up contact with the respondent after the roster is set. (Butts, 2008; Heckathorn, 1997).

2.3.2 Ego-Network Studies. In ego-network studies (Jones & Volpe, 2011), the network data is collected with respect to an individual subject. The "ego" refers to that studied subject to whom the researchers ask to enumerate the alters (other subjects within the social network) that they recall and report their relations with these alters. In these studies, respondents may also be asked to report ties among these alters. This method is mostly employed when the network is assumed to be large, and its members are not known prior to the study. Respondents are asked

2.3.3 Cognitive Social Slices. Cognitive Social Slices (CSS) is a technique intended to predict the distortion that occurs due to informant inaccuracy. In what follows, we summarize Krackhardt's explanation of CSS (Krackhardt, 1987).

In the CSS survey, respondents are asked about their relations between others and also their perception of ties between alters. Their responses are stored in a $N \times N \times N$ network matrix. In this matrix each cell is denoted as $R_{i,j,k}$ where i is the sender of the relation, j is the receiver, and k is the perceiver of the relation (Krackhardt, 1987).

After the data is collected from respondents, all reports are aggregated and two different data slices are generated, namely the Locally Aggregated Slice (LAS) and the Consensus Structure Slice (CSS). In LAS, responses are aggregated according to the intersection rule or the union rule. In the union rule, if at least one of the endpoints of an edge reports a tie, the tie is included in the LAS. In the intersection rule, both endpoints' must report the existence of a tie in order to ensure its inclusion. When we construct the CSS, a threshold value is used, and a link is included in the CSS structure if and only if the number of reports which perceive the existence of the tie is equal to or greater than this threshold value.

2.3.4 Response Format. After discussing the survey and interview methods, we will briefly describe the response format that is used to encode relational network data. When collecting relational data, respondents may be asked to report their judgments and perceptions in a binary form, ordinal form or as a ranking (Marsden, 2005; Wasserman & Faust, 1994a).

Binary encoding is the simplest method for respondents. In this approach, respondents simply indicate whether there is a tie or not (Marsden, 2005). For example, respondents may be given an enumerated list or roster and asked to check each of the ties they perceive, or respondents may be asked to generate a list of the

names of people with whom they believe they have ties (Coleman, Katz, & Menzel, 1957; Parker & Asher, 1993; Marsden, 1990).

In comparison, the ordinal rating or ranking formats require that respondents evaluate their relationship strength, weight, frequency. In ordinal rating, In ordinal rating, respondents are asked to rate all the other actors for a particular measure. In ranking format respondents rank their ties to all other actors within the social network. (Bernard, Killworth, & Sailer, 1979; Wasserman & Faust, 1994a; Jones & Volpe, 2011).

2.3.5 Potential Shortcomings. Because we will be designing and presenting a new technique to collect relational data, it will be useful to evaluate the shortcomings of the existing schemes described above.

Whole Network/Roster Method, Cognitive Slices The researcher provides a list of all pairs of subjects in the sample, and asks respondents to report the ties between them. Unfortunately, if the number of people in the sample is very large, listing all members in the roster becomes impractical, so the researcher must select a subset from the sample and asks the respondent to report ties within this subset.

Ego Network/ Free Recall Method Respondents enumerate individuals in their ego-network and report their perceptions of existing ties.

Unfortunately, these schemes suffer from several shortcomings: inefficiency, distortion and concerns of privacy and anonymity. We will elucidate the nature of these concerns by way of a small example.

Example: Suppose that we are studying a very small network illustrated in Figure 3 where the adjacency matrix is given in Table 1.

There are eight actors in the network. Suppose that six of the actors are in the sample whereas actor C and G are not included in the sample. Then, respondents are provided a roster or a list in which these six actors are enumerated.

If we assume all ties are disclosed by respondents accurately,(which is the most optimistic situation) methods such as roster or cognitive slices inevitably lead to inefficiency and distortion in the collected network data.

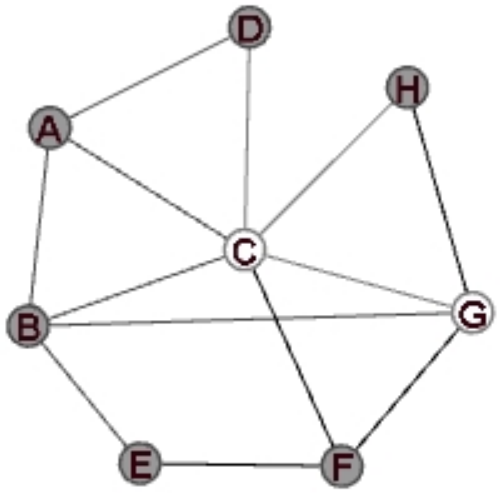


Figure 3. A graph with 8 nodes, where 6 nodes are included in the sample.

	A	B	C	D	E	F	G	H
A	0	1	1	1	0	0	0	0
B	1	0	1	0	1	0	1	0
C	1	1	0	1	0	1	1	1
D	1	0	1	0	0	0	0	0
E	0	1	0	0	0	0	1	0
F	0	0	1	0	1	0	1	0
G	0	1	1	0	0	1	0	1
H	0	0	1	0	0	0	1	0

Table 1
Adjecancy matrix

a. Inefficiency: This method is unable to retrieve all available information. Suppose that respondents know members C, D and H where C, D, H forms a geodesic between D and H. If C is not included in the roster (sample), respondents can not report the ties between D and C , and C and H. Therefore, the shortest distance between D and H is not revealed, even though respondents know this information.

Another inefficiency may occur if the network is sparse. Suppose we have a large community where members have very few links between them. In such a circumstance, asking subjects to identify ties between a subset of samples may reveal very few ties, which results in inefficient use of interviewing resources.

b. Distortion in distances: Geodesics that are produced from this method will suffer from distortions because shorter paths may exist which are not discovered from the sample data. For example, in estimating the geodesic distance between actors D and F, a longer path, (D,A,B,E,F) may be discovered by chance because of the non-discovery of an intermediary node, for example C.

Also note that actor H appears as an isolated node because actors C and G failed to be subjects in the study. Thus, there can be the distortions in all pairs of geodesic distances between the original graph and estimated graph since the estimated graph only includes links among actors in the sample.

	A	B	C	D	E	F	G	H
A	0	1	1	1	2	2	2	2
B	1	0	1	2	1	2	1	2
C	1	1	0	1	1	1	2	1
D	1	2	1	0	2	2	3	2
E	2	1	1	2	0	1	2	1
F	2	2	1	2	1	0	3	2
G	2	1	2	3	2	3	0	1
H	2	2	1	2	1	2	1	0

(a) Original graph

	A	B	D	E	F	H
A	0	1	1	2	3	∞
B	1	0	2	1	2	∞
D	1	2	0	3	4	∞
E	2	1	3	0	1	∞
F	3	2	4	1	0	∞
H	∞	∞	∞	∞	∞	0

(b) Predicted graph

Figure 4. Geodesic distances between all pairs in a graph of 8 nodes

If we now turn to ego-network methods, these methods suffer from privacy and anonymity concerns.

c. Privacy: Respondents may be very sensitive about their privacy because of criminal or stigmatized characteristics of the social network. Therefore, when they are asked about their connections with others, they may feel their privacy is violated, and as a result, may produce false or biased reports (Dombrowski, 2012).

Suppose that B is the respondent, then if B denies its ties with A and E , the true distance between A and E may not be discovered. As explained above, this will cause a data loss and associated distortion.

d. Anonymity: In the free-recall method, respondents may not want to cooperate with the researcher in listing their alters because they assume it will jeopardize their anonymity. Respondents may not wish to disclose such a list because of an implict social norm against "snitching" (Heckathorn, 1997). If we turn to our example, they may not want to report links where actors C and G are involved, because this requires nominating people who are not in the sample.

These problems above a thorough d can not be overcome by increasing the sample size, because researchers have limited resources (e.g. incentive money and interviewing time) (Arsovska, 2012).

3 Problem Statement, Objectives and Assumptions

Problem Statement: Is it possible to estimate topological features of a social network by asking a sample of individuals about the perceived proximity of *other pairs* of individuals? If so, what are the factors that influence the performance of such a scheme?

This scheme must address the following important objectives:

- **Efficiency:** The scheme should handle sampling and data collection in such a way that the network topology will be revealed with a minimal amount of resources (e.g. time and incentive money).

- **Accuracy:** The estimate of topological features produced from the collected data should indicate closely resemble the feature of the the actual network.

- **Anonymity:** The data collection mechanism should allow the respondents to preserve their anonymity, and safeguard against reported relations being trace

backed to the studied subject.

- **Privacy:** Because informants are reluctant to disclose their own ties, the scheme should favor asking them about the proximity between other pairs of individuals.

Approach: The core idea of this research is a new model for interview design which asks the respondents to report their perception of proximity between other pairs of individuals in their social networks.

We have two assumptions in this approach:

Assumption 1: The main assumption of our study is that people's perceptions of social proximity between recognized alters is coherent with the geodesic distance between those alters.In other words, respondents are expected to have significantly different perceptions of proximity for pairs of individuals that are at different geodesic distances from each other. For example, a respondent reports about three different pairs, who are at geodesic distance one, two and three. We assume that the respondent will perceive and report smallest distance for the first pair and the greatest distance for the last pair.

Assumption 2: When the geodesic between a pair is very large, respondents can not make an assessment about the social proximity between this pair. We use "perceivable proximity threshold" to refer the range of distance where respondents can distinguish distances among members in a pair.

4 Approach

In this section, we will first introduce a new interview design method that may be used to collect information about large social networks and hidden populations. Then we will describe our simulation environment with which the interview design was evaluated and optimized.

4.1 A New Data Collection Method

We designed a new data collection method for social networks, that may be used when:

a. No network information is available beforehand.

b. Respondents are sensitive about their privacy and anonymity. In particular, respondents do not want to enumerate the alters, or to disclose their own ties.

Assuming these characteristics hold in the network of interest, the researcher samples individuals, and conducts interviews with them. After completing each interview, the researcher records the subject's photograph. The ongoing process yields a set of pictures of subjects which grows by one upon the completion of each interview.

The interview consists of two phases.

In the **marking phase**, respondents are shown the icons, figures or names (any symbols which the actors are well-known with and helps the respondents to recall associated subject) from the sample, and they are asked to mark those subjects that they recognize, and separate from them the ones that are unfamiliar to them. In this phase, respondents may also be asked to report the social proximity between them and the recognized subjects in the list.

In the **prediction phase**, respondents are shown pairs of recognized subjects and asked to report on the perceived proximity (1,2 or 3) between each pair. We assume that the perceived proximity that is reported is proportional to the geodesic distance between the pair within the social network. For example, if a respondent perceives that a pair of individuals reciprocally know each other, the respondent will report 1 as the perceived pairwise distance. On the other hand, if the respondent believes that the two subjects have a mutual friend, but do not know each other directly, the respondent will report 2. Finally, if the respondent believes

that one member of the pair has a peer who is acquainted with a peer of the other member of the pair, the respondent will report 3.

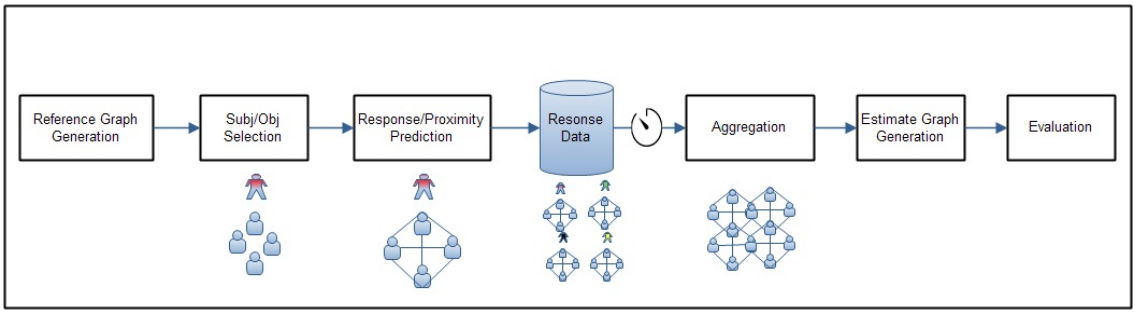


Figure 5. SNAPT system overview

4.2 Overview of Simulation Environment

In order to test the performance of our data collection method, we developed a simulation environment which enables us to conduct experiments with different sampling methods.

Figure 5 shows the core model for a SNAPT simulation environment. The simulation, in brief, follows these steps:

Step 1: We generate a reference graph which plays the role of the true social network, with all of its attributes: members (vertices), social ties (edges), and social proximity (shortest path distance).

Step 2: We simulate the respondent recruitment by selecting a vertex from the reference graph according to two sampling methods: random sampling and respondent-driven sampling.

Step 3: After a respondent has been chosen, an object selection process takes place. In object selection, the goal is to select a subset of the sample, for which respondents could efficiently estimate social distances. To achieve this goal, we develop three different selection algorithms: random sampling, sampling from recognized objects, and sampling first from perceivable proximity.

Subject	Vertex	Vertex	Proximity
1	2	3	2
4	2	3	2
3	1	4	3

Table 2
Pairwise distance in response data

Step 4: After selecting the object set, we simulate respondents' estimates of social proximity between object pairs.

Step 5: According to the proximity prediction scheme, we select the proximity values which are within the assumed perceivable proximity threshold, denoted as pm. The proximity reports are denoted as $R_{i:j,k}$ where i is the respondent, j and k are the objects whose pairwise distance is being reported, is included into response data according to the formula below:

$$R_{i;j,k} = \begin{cases} R_{i;j,k}, & \text{if } R_{i;j,k} \leq \text{pm} \\ \varnothing, & \text{otherwise} \end{cases}$$

These reports are then stored in a data structure which is shown in Table 2.

We repeat Step 1 through 5 until a pre-determined number of interviews have been completed.

When the number of completed interviews reaches the pre-determined number, we evaluate the model's performance by aggregating the data collected up to that time. To accomplish this, we followed steps 6 through 8.

Step 6: We aggregate the information in the response data in two steps:

First, for each distinct pair, the average value for all respondent reports is calculated; we refer to these as "Type-I Proxy" distances. Second, pairwise distances that are not reported directly are inferred using the reported distances. For example Let $d_{j,k}$ be the distance between two vertices j and k. If $d_{A,B}$ and $d_{A,C}$ are reported by some subject, we are able to predict the distance $d_{B,C}$ even if it has not been

reported by any subject by traversing the path $d_{B,A}$ and $d_{A,C}$. In this way, we can derive estimates of pairwise distances even if no respondents have reported the distance for that pair.

Step 7: After aggregation, we generate an estimated network based on the aggregated distance information.

Step 8: To evaluate the accuracy of the estimated network, the reference network and the estimated network are compared using a variety of algorithms we developed: all pairs correlation, routed correlation, node discovery rate. Finally, results were reported in proper graph formats.

We repeat steps 1 through Step 8 until the desired number of interviews has been completed.

Step 9: When the simulation has completed the required number of interviews performance reports are generated in different formats.

A single run of the simulation is illustrated in Figure 6

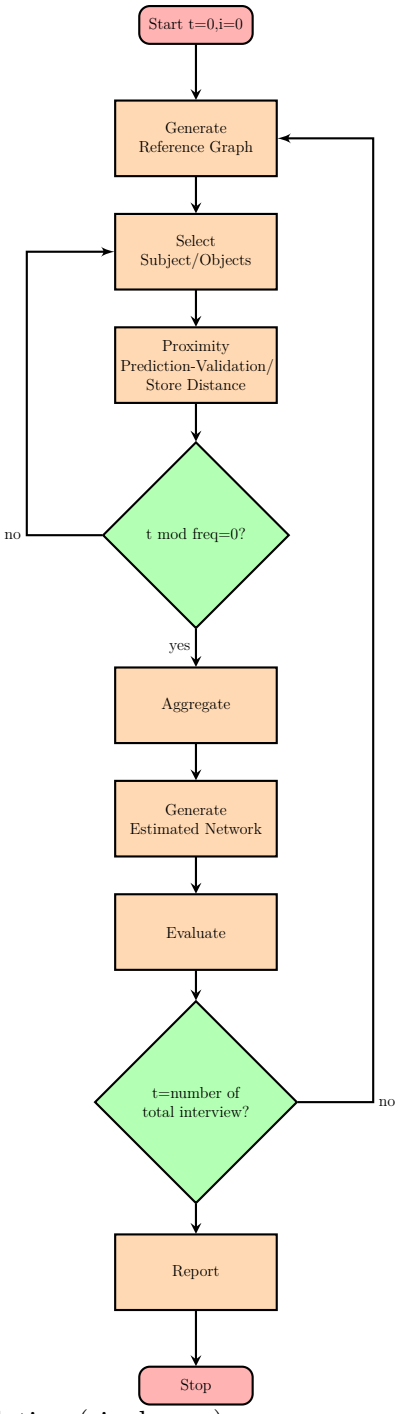


Figure 6. Flowchart for simulation (single run)

In the following part of this section we will present each steps in detail.

4.3 Generating True/Reference Graph

The characteristics of the reference graph are described below:

The reference graph, denoted as $G_r(V, E)$, is generated as an undirected graph with vertices, V, and edges, E.

Vertex and Edges:

Each vertex $\in V$, has an ID number and owner network attributes. The ID attribute is a unique number which identifies the vertex in the graph, and the network ID is an attribute used for identifying the network to which the vertex belongs.

An edge is denoted as $e(u, v, w) \in E$ where u , $v \in V$ are the endpoints of the edge, and w is the weight of the edge. In our experiments, the reference graph is generated as an unweighted graph, and for this reason w was always taken equal to 1.

We represent these attributes in the adjacency matrix A = $(a_{i,j})$ for $i = 1, ..., N$ and $j = 1, ..., N$.

$$a_{i,j} = \begin{cases} 1, & e(i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

Given these representation of a graph, we now present how we generate the graph.

Method for Generating Reference Graph

In order to generate the reference graph, we implement a Barabasi-Albert model, which is a probabilistic scheme to control the growth and link attachment process for scale-free networks (Barabasi & Albert, 1999). We select the Barabasi-Albert model, because it is simple to implement, and it is a realistic model

that has been validated by numerous researchers, and shown to reflect the topology and degree distribution of the actual social networks (Schneeberger et al., 2004; Dombrowski, Curtis, Friedman, & Khan, 2013).

In the Barabasi-Albert model,at each step, one vertex is added to the graph, and this vertex is attached to a number of existing vertices according to the "preferential attachment probability" (Barabasi & Albert, 1999).

In order to generate a Barabasi-Albert graph, the following parameters must be specified:

- Number of Nodes, N: The number of vertices in the graph, which represents the number of members in the SN.
- Number of Initial Nodes, m_0.
- Number of edges to attach, m: This parameter determines the number of vertices to which each new vertex attaches when it joins to a network. Note that m must be less than equal to m_0. This parameter determines the growth in the number of edges in the Barabasi-Albert model.

Given these parameters, the preferential attachment probability for a vertex u, is calculated with using the equation below:

$$\rho(u) = \frac{k_u}{\sum\limits_v k_v} \text{ where } u \neq v.$$

Here, the enumerator is the degree of vertex u and the denominator is the sum of degrees of all vertices. In this model, vertices that have higher degrees are more likely to be assigned edges to newly created vertices.

Implementation

At the beginning the graph generations, a new graph is created with m_0 isolated vertices. The vertices are assigned sequential ID numbers.

After the initial nodes are added in the graph, new nodes are added one at a time with sequential IDs until the size of the network reaches N. Whenever a new

vertex is added in the graph, m vertices are selected, and edges are added between the new vertex and these m vertices. The m are selected according to the preferential attachment function given below.

$$\rho(u) = \frac{k_u + 1}{|E| + |V| - 1}$$

At the implementation layer, this is achieved using a random number generator, which produces a real number between 0 to 1. A vertex is selected randomly from the previously vertices. Then preferential attachment probability is calculated for this vertex. If the preferential attachment probability is equal or greater than the random number, an edge is established between the new vertex and the selected vertex. Otherwise, another vertex is selected randomly, and these steps are repeated until the preferential attachment value of the selected vertex is greater than the random number.

The pseudocode for generating the reference graph is presented below.

Algorithm 1 Pseudocode for generating Barabasi-Albert graph

```
Initialize Graph (m_0)
for i = m_0 → N do
    Create a new Node, u
    for j = 0 → m do
        Select a Random vertex, v, from the graph
        Calculate preferential probability to attach for vertex, v,
        Generate a Random Number between 0 and 1, p, for this vertex
        if preferential probability ≥ p then
            Create Edge, e(u, v)
        end if
    end for
end for
```

Custom Java classes that we develop are presented in Appendix A

Figure 7 shows a reference graph of 500 nodes.

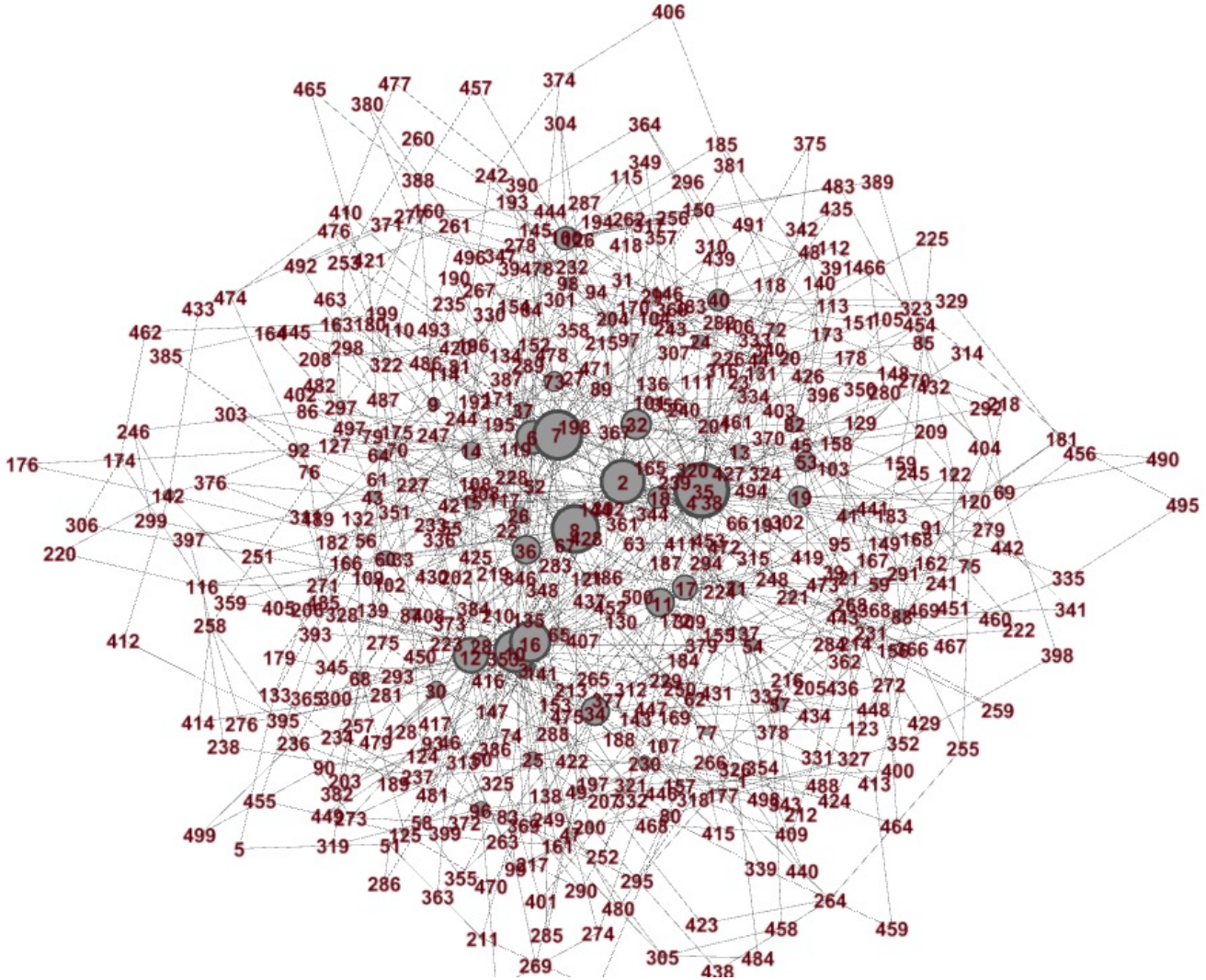


Figure 7. The graph visualizes a 500-nodes graph which is generated implementing the Barabasi-Albert model. The circles are the vertices, and the numbers are unique ID numbers corresponding to these vertices. The sizes of vertices are adjusted in proportion to the node degree.

4.4 Subject Sampling

We use the word "subject" to refer to the respondents who participate our study and report their perception about alters' social distance. In this research, we test two subject selection algorithms: random sampling and the respondent-driven sampling. In each method, the system selects one vertex at a time from the reference graph, and simulates the process of recruitment.

4.4.1 Random Sampling. Random sampling is a probabilistic method in which each subject is assigned an equal probability of being selected for the study period. We implement two types of random sampling methods: sampling with replacement and sampling without replacement.

- Random Sampling with Replacement: In this selection method, we allow a subject to participate in the survey more than once. Thus, for each subject selection, the probability for an individual v to be chosen is taken to be

$$\rho(v) = \frac{1}{N}$$

and the probability for a vertex to be selected at least once by time t is therefore

$$\rho(v >= 1) = 1 - (1 - \frac{1}{N})^t$$

Note that one subject is selected at a time, and the probability changes according to t and N. Here t refers to the total number of interviews that have been conducted so far. The formula implies that the greater the N the smaller the probability of selection.

- Random Sampling without Replacement: This method does not allow a subject to participate in the study more than once. The probability for a

subject to be selected for participation in the study at time t is

$$P(v) = \frac{1}{N - t}$$

. The probability for an individual to be selected at least one time at time t is

$$\rho(v >= 1) = \frac{t}{N}$$

.

Whenever we do not have prior knowledge about the research population, sampling without replacement yields a higher number of unique subjects compared to sampling with replacement.

4.4.2 Respondent Driven Subject Sampling. We simulate the RDS recruitment process as follows. First, an arbitrary number of vertices are selected randomly from the reference graph. These subjects stand for the seeds, and receive the first set of the RDS coupons. We use the phrase "coupon source" to refer to the subjects have been given RDS coupons.

New subjects are selected randomly from the peers of previous respondents (Salganik & Heckathorn, 2004). In other words, recruits are chosen from the vertices who are at distance one from the coupon source which still have residual credits. Whenever a coupon is transferred, the number of coupons associated with the respondent is decreased by one, and the new subject is given three coupons.

Let S be the subjects who participate in the research, we select new subjects and track the RDS coupons with the below algorithm:

Algorithm 2 RDS subject sampling and coupon tracking

Step 1. Filter the subjects in S who have remaining coupons and have any neighbor (distance=1) that has not participated in the interview.
Step 2. Put the filtered subjects into a list.If the list is empty we are at the maximum number of subjects that can be reached by RDS, so then we exit.
Step 3. Select a random subject ,coupon source, from this list.
Step 4. Visit all neighbors of the new coupon source who have not already participated in the interview, and store these neighbors in a candidate list. If there is no such neighbor (i.e. the candidate list is empty), go back to step 3.
Step 5. Select a new respondent from the candidate list.
Step 6. Give the new respondent three coupons and add it to S. Return to Step 1.

4.5 Object Selection

We use the term "object selection" to refer to the process of efficiently choosing a set of previous subjects to show the current respondent, and ask them about perceived pairwise social distances.

Here we will present different object selection methods for choosing objects from the set of studied subjects so far. Initially, the set of studied subject is empty. After each interview, the set of studied subjects expands by one.

Henceforth we refer to the subject database as S.

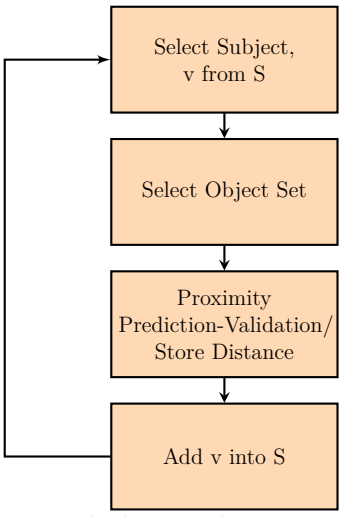


Figure 8. Flowchart for the subject and object selection

General Characteristics of Sampling From a Live Data Set:

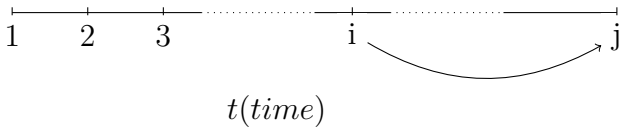


Figure 9. Snapshot of S when t=j

a. S grows one subject at a time, the addition of a new subject into set S results in $t-1$ new pairwise distances which may be perceived by future subjects.

b. Let t be the current time of an interview, there are $\binom{t}{2}$ pair of subjects for which distance may be perceived.

c. Let k be the number of objects selected to be shown to the subject in the t^{th} interview. At time t, the ratio of pairs that have already been shown to the total number pairs that are available is approximately equal to $\frac{t\times\binom{k}{2}}{\binom{t}{2}}$.

d. The new subject is added to the set S, and becomes a candidate to be selected as an object that will be shown to the future subjects.

This object selection process was varied by considering different selection algorithms: sampling with replacement, sampling from recognized objects and sampling within ego network first. In the next section, these algorithms are discussed in detail.

4.5.1 Random Sampling with Replacement. In this method, for each interviewed subject, k vertices are selected randomly with replacement from S. The probability for an individual v to be selected as an object is uniformly given by

$$\rho(v) = \frac{k}{|S|}$$

.

This method suffers from a bias. The reason for this is that all elements in S have the same probability of being selected, so the expected number of times that an element of S is chosen to be an object is greater for older members of S. In

particular, if a subject v is interviewed at time $t = j$, the probability of being selected is for vertex v:

$$p(v) = 1 - (1 - \frac{1}{n})^k$$

It follows that the expected number of times that subject v will be shown as an object is

$$Expected(v) = \sum_{n=j}^{n=t} 1 - (1 - \frac{1}{n})^k$$

.

4.5.2 Sampling within Recognized Subjects. In this method, we assume that a subject is able to recall only a subset of the larger community. We use a parameter which we can the "recognition number" to specify the number of people a subject is able to recall.

The recognition number will be denoted as rn. Let N be our estimate of the total population size, and let fixed rn be the recognition number. Using N and rn, we can calculate the recognition ratio,which is denoted as r, which determines how likely a subject is to recall a randomly chosen individual from the social network:

$$\mathrm{r} = \frac{\mathrm{rn}}{\mathrm{N}}$$

For a set of subjects, S, that is growing by one with each consecutive interview, the number of people that the subject who is interviewed at time t will recall is

$$K = r \times (t - 1)$$

For example, if the 51^{st} subject who participates in the interview knows 100

people in the social network, and this social network consists of 1000 people, then our estimate for K is $50 \times \frac{100}{1000} = 5$. This means that at the time we interview the 51^{st} of 1000 subjects we expect this subject to be able to recognize 5 of the 50 subjects we have interviewed previously.

In our implementation, each time we attempt to select K objects to show the current interviewee, we must calculate K and determine the number of objects that are to be selected. Note that the system can not select any objects until the estimate of K is greater than or equal to 1, and K increases linearly in proportion to the number of subjects interviewed until it reaches its maximal value which we take to be the recognition number(rn).

$$0 \leq K \leq \text{recognition number}$$

There are three parameters of these methods:

- rn: Recognition Number. This is the maximum number of individuals that a subject is expected to be able to recognize.
- S: An ordered list of distinct subjects who have participated in this survey so far.
- N: An integer value which is an estimate of the size of the total population

Given the above parameters, the pseudocode of the method is presented below.

Algorithm 3 Pseudocode for selector

$K = \lfloor(\frac{|S| \times rn}{N})$
if $K \geq 1$ and $K \leq rn$ **then**
  Select K objects randomly from S
end if
Add new subject to S

4.5.3 Sample Within Ego Network First (ENF). In this selection method, we assume that there is a perceivable proximity threshold,denoted as pm, and respondents can estimate the distance between themselves and other subjects or between pairs of subjects in the community as long as the distances are less than or equal to the perceivable proximity threshold.

Within each interview a respondent is asked to separate the known subjects that have been seen so far into two categories: those that they recognize and those they do not recognize. Then the subject is asked to estimate the distance between them and the subjects they have recognized. We choose pairs of objects that are within the perceivable proximity threshold distance from the interviewee.

A Class diagram where the custom classes used to implements can be seen in Appendix A.

4.6 Distance Calculation/Proximity Prediction

After object selection, all pairs of the shortest path distances among object set, denoted O, are calculated using Dijkstra algorithm. Thus, the number of pairwise distances computed is $\binom{K}{2}$.

The Dijkstra algorithm is an algorithm to compute the shortest path between a given source node and destination vertex.

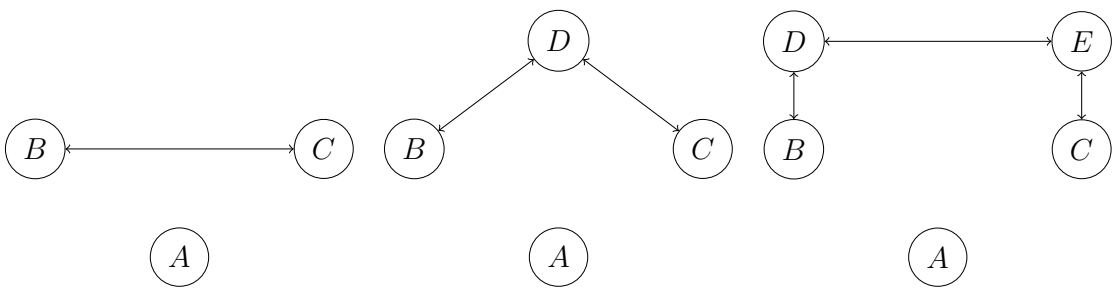
4.6.1 Infinite Perception. In this model, we assume that subjects can produce perfect estimates about the proximity of pairs of subjects in the studied network. In other words, a subject when queried about the distance between any two previously known subjects will be able to provide an accurate estimate no matter what the separation between these two vertices is.

4.6.2 Validation Perceivable Proximity. In this model, we assume that;

1. A subject can estimate the distance between a pair of vertices, if and only if the true distance between these pair of vertices is less than a certain perceivable perception threshold.

2. The respondent perception reports which are greater than the perceivable proximity threshold is inaccurate, so they need to be omitted.

We illustrate how the perceivable proximity threshold operates by way of an example.



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Figure 10. In this figure three scenarios for the distance between B and C are illustrated.

Suppose that $pm = 2$, and A, B, C, D and E are people in the social network, and their positions are illustrated as shown in Figure 10.

This model assumes that A may reliably report an estimate of social proximity between B and C, only if B and C knows each other, or they have a mutual friend, and do not know each other. Thus, A can report the proximity between B and C in Scenario 1 and 2. However, in Scenario 3, A's perception will be inaccurate because the proximity between B and C is 3 which is greater than pm.

In the implementation, we calculate the distance between object pairs, and store the distances which are less than or equal to the perceivable proximity. We omit the distances which are greater than the perceivable proximity.

4.7 Aggregation

In our simulation for a distinct vertex pair, several reports may be given by different subjects. We merge all these reports into a single record by calculating average values. We called these distances Type-I proxies. In other words, for any reports which involve the same vertex pair, we calculate the mean value of all the reported distances for that pair, and use this mean value as our aggregated estimate.

In addition, for all pairs whose distance has not been reported directly by some subjects, we determine whether there is a path between the endpoints of the pair. If such a path exists, connecting the pair we use the length of shortest such path as an estimate of the distance. If no such path exists, we record the length for this pair as infinity.

We illustrate this aggregation process by way of an example using a five node graph shown in Figure 11. Suppose that the response data before aggregation is illustrated in Table 3.

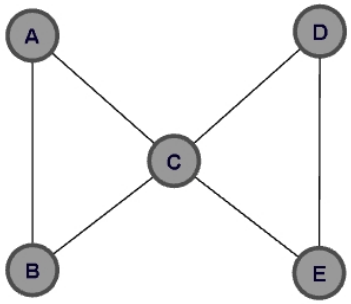


Figure 11. An example graph of 5 nodes

Subject	Vertex	Vertex	Proximity
D	A	B	1
C	A	B	1
B	A	D	2

Table 3
Response data(Before aggregation)

First, the responses for pair (A, B) are merged. Since both C and D report the distance between A and B as 1, the average taken to be the canonical estimated distance which is 1. The distance between B and D is not reported directly, however, the path which involves links (B,A) and (A, D) can be found. Because no other path can be distance. The aggregated data is illustrated in Table 4.

Vertex	Vertex	Proximity
A	B	1
A	D	2
B	D	3

Table 4
Aggregated data

4.8 Estimated Graph

The estimated graph is generated as a cumulative representation of the responses using the results from the aggregation process decribed in the previous section.

The estimated graph is generated as a weighted graph, denoted as $G_{Est}V, E$

where the vertices in the aggregated data are added as nodes, and both Type-I and Type-II proxies are added as edges.

The custom classes used in this method of generating an estimated graph is presented in Appendix A.

4.9 Evaluation

After generating the estimated network, we compare it to the reference graph and measure the divergence between them.

In the estimated graph, the distance between two vertices u and v can be;

a. finite, implying that there is a path connecting u and v.

b. undefined, meaning that at least one of the vertices are not included in the estimated network,

c. is infinite because there is no path between u and v.

Pairs of vertices which fall into the first of these categories, provide numbers which can be correlated with the ground truth as measured using the reference network. In other words, we can correlate pairwise distance values in the estimated network against corresponding distance values in the reference network. The pairs which lie in the second group are used to give rise to the notion of node discovery. The pairs of vertices which fall in to the third group are not included in correlation computation because these values are infinite.

The performance metrics we use include: vertex discovery rate, adjusted vertex discovery rate, all pairs distance correlation and routed correlation.

4.9.1 Vertex Discovery Rate. Vertex discovery rate shows the ratio between the number of vertices in the estimated network and the total number of subjects in the set S.

4.9.2 All Pairs Distance Correlation. In all pairs correlation, we calculate all pairs distance in the estimated network and all pairs distance in the

reference network. Then, for pairs (u,v) of vertices of which these two distances are finite, we calculate Pearson Correlation as follows.

$$\rho(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}}.$$

where X is the set of all finite pairwise distances in the estimated graph, and Y is all corresponding distances in the reference graph.

4.9.3 Routed Correlation. In routed correlation, we implement a vertex-centric correlation measure to evaluate accuracy of the estimated network.

Let V_{Est} be the set of vertices in the estimated graph. For any vertex v $\in V$, we calculate the distance to all other vertices. This results in $N-1$ pairwise distance from v to the other nodes in an N-node network. Next, we correlate these numbers with their corresponding distance values in the reference graph.

After calculating vertex-centric correlation coefficients for each vertex v $\in V$, we obtain a set of N correlation values. In order to predict overall performance, we calculate the mean of all vertex-centric correlation values.

The pseudocode for Routed Correlation is presented below:

Algorithm 4 Routed correlation coefficient

```
for i = 1 → N do
  src = v_i
  for j = 1 → N do
    if est_D(src, v_j) ≠ ∞ then
      P_x[j] = d_e(src, v_j)
      P_y[j] = d_t(src, v_j)
    end if
  end for
  R_i = ρ(P_x, P_y)
end for
RoutedR = mean(R)
```

4.9.4 Adjusted Discovered Nodes. In this method, we address some of the shortcomings of our correlation coefficient calculations. Our correlation

coefficient calculation methods may report an inappropriately high correlation coefficient when there are relatively few number pairs of distances of finite-distance pairs and many pairs of vertices at infinite-distance pairs.

In the adjusted discovered nodes method, the fraction of pairs which are finite distance is used to correct this type of error. Let D be the number of nodes in the estimated network. We can calculate the adjusted discovered node with below formula:

$$\text{Adjusted Discovered Nodes} = D \times \frac{\text{Number of finite pairs}}{\binom{D}{2}}$$

The custom classes used for evaluation are presented in Appendix A

4.10 Reporting

In this part, we present the structure of report files which are produced in the course of simulation.

4.10.1 Text File Format. For each run of the simulation, a report file is produced by the system. The format of the text file is presented in table below.

Interview Number	Number of Nodes Discovered	Number of Pairs Compared	Correlation Coefficient

Table 5
Text file format reporting a single experiment

After the simulation completes all runs, the system produces as many reports as experimental trials in text file format. Then using a code script, corresponding values of these text files are read, and the mean and standard deviation information are calculated and reported in a text file. The format of this file is presented in table below.

Interview Number	Mean of Correlation	Std. Deviation of Correlation	Mean of Vertex Discovery	Std. Deviation of Vertex Discovery

Table 6
Text file format reporting compiled results

4.10.2 Dot File Format. In order to visualize the reference and the estimated networks, the social network data is written to a graphics (.dot) file (Gansner, Koutsofios, & S., 2006).

5 System Development

For this study, we developed a custom simulation environment using Netbeans IDE with Java 1.7.0.4 for Window 64 bit (Oracle, 2014).

In our software, in addition to standard Java packages we also used Apache's open source mathematical library, Apache Commons Math (*Apache Comman Math Library*, 2009).

The source code of the simulation environment can be found at the git repository located at $http://git.code.sf.net/p/snapt/code$

With regards to software design, we aim to develop the software in a modular way where main interfaces are created for each component such as network, selector, divergence calculator etc. Thus, the program can use classes which implement base interfaces without requiring significant changes to the source code. It also enables the addition of new implementation classes to the the system.

In input for our software is a text file where the following parameters are defined:

- the seed of the random number generator,
- parameters of the reference graph: network size, number of initial,nodes and number of edges to attach at each state of Barabasi-Albert network generation,
- Selector Type: Random, RDS, RDS-Ego,
- Evaluation Metric: All Pairs Correlation, Routed Correlation,
- Recognition Number,
- Perceivable Proximity Threshold,
- Number of subject interviews (simulation termination criteria)
- Frequency to get evaluation results

When the simulation is run, the above input file is parsed, and the system initializes classes and parameters using the fields specified in the input file.

After this initialization, the simulation runs as shown in the Figure 12.

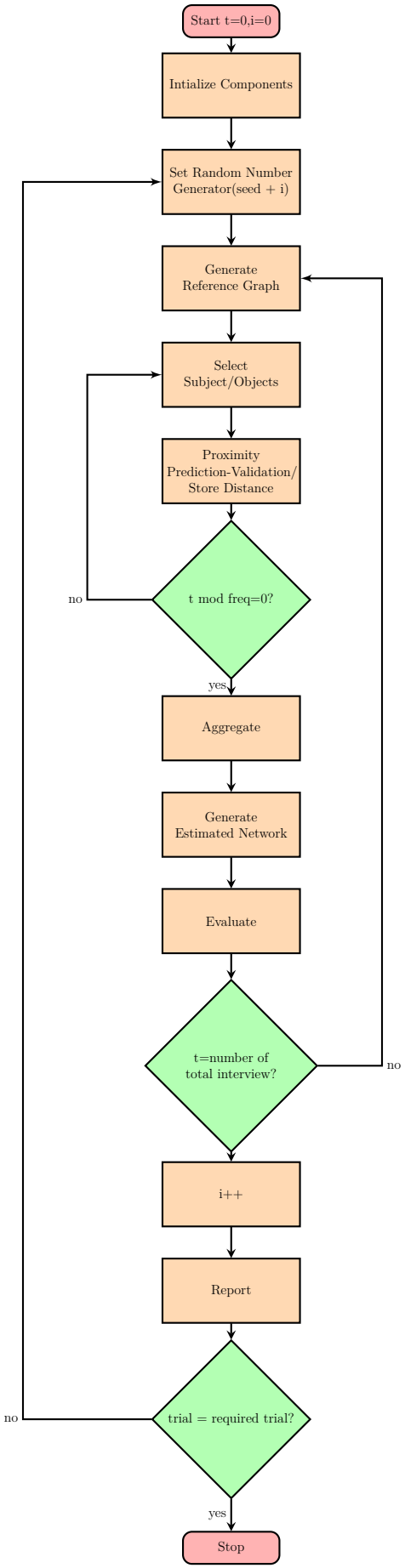


Figure 12. Flowchart for SNAPT simulation

	Model-1	Model-2	Model-3
Subject Selection	Random Sampling without Replacement (pg.29)	Respondent Driven Sampling (pg.26)	Respondent Driven Sampling
Object Selection	Random Sampling from Recognized Subjects	Random Sampling from Recognized Subjects	Sample First From Ego Network (pg.31)

Table 7
The subject and object selection methods implemented in models

The class diagrams for the system can be seen in Appendix A.

6 Experiments

In this section, we will first present three models that we have developed using different combinations of subject selection and object selections. Then, we will present test results which compare the performance of these three models.

Table 7 lists the three models and the subject/object selection techniques that are used in each of them.

In Model-1, the subject selection is executed as random sampling without replacement from the total population, and the object selection is performed randomly from among recognized subjects who have been recruited previously.

In Model-2, subject selection uses Respondent Driven Sampling while object selection is performed the same model as for Model-1.

In Model-3, respondents are not only asked to identify subjects that they recognize, but also they are required to estimate the proximity between themselves and the subjects they recognized. In this model, the subject selection is performed using RDS, and object selection is using the Sample Within Ego Network First method.

In all three models above, the objects that are shown to a subject are chosen from those subjects who are recognized among previous recruits. Therefore, the number of subjects recognized, denoted as K, is calculated on-the-fly for each

interview based on the recognition number on the discovered network size.

In each of these models, the experiments are performed on reference networks of 1,000 and 10,000 nodes which were generated according to Barabasi-Albert model.

The reference graphs are generated with the following parameters.

- Number of Initial Nodes of Barabasi-Albert Model, m_0= 10,
- Number of edges to attach for each new vertex, m=2.

All of the experiments are performed assuming two different perceivable proximity thresholds. In one set of the experiments pm=2, and in another set of experiments pm=3.

The experiments are conducted for rn 100,200,300 or 1000.

To evaluate the performance of each of the three models, we calculate all pairs distance correlations, and report the adjusted vertex discovery numbers during the course of interviews. Each experiment is conducted ten times and the results of these ten trials are given along with error bars to show mean and the standard deviation of performance outcomes.

The experiments are tested on computers which had 64 bit Intel I7-2600, 3.6Ghz CPU and 16 Gb. RAM. The operating system is Windows 7 Enterprise with Service Pack-1.

6.1 Model-1 - Random

In this model, subjects were selected from the reference graph randomly without replacement. After subject selection, the objects are selected from previously interviewed subjects as follows.

First, the number of objects that the subject can recognize is calculated via below equation:

$$K = \frac{\text{Recognition Number}}{\text{Network Size}} \times (t-1)$$

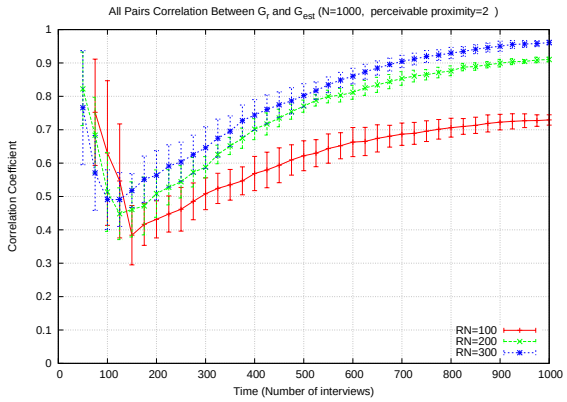
where t is the number of interviews completed. Then K objects are selected randomly among previously interviewed subjects. After object vertices have been selected, proximity prediction is performed.

After K objects have been selected, they are shown to the subject, and the subject attempts to estimate the pairwise proximity between these K objects. For each of the $\binom{K}{2}$ pairs, if the pairwise distance was less than or equal to the perceivable proximity threshold, the subject is assumed to provide an accurate estimate of the distance between the two vertices. If, on the other hand, the pairwise distances which are greater than the perceivable proximity threshold, then the subject is unable to provide a distance estimate for the pair.

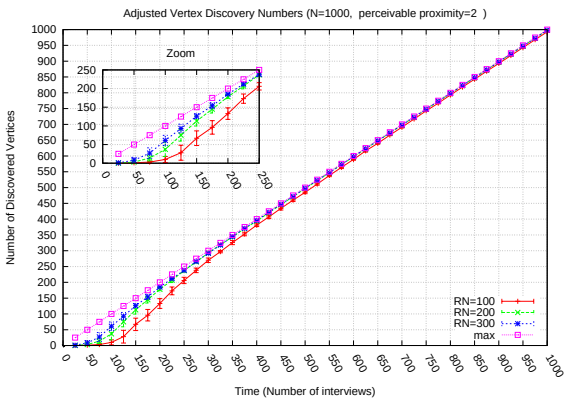
After every 25 interviews, we take a snapshot of the system, run the aggregation process, estimate the resulting network, and carry out the evaluation of the accuracy of the estimated network.

In greater depth, firstly the response data was aggregated (see pg.33), then an estimated network (see pg. 34)is computed, then this estimated network is used to compute all pairs shortest path distance (see pg. 35), then the all pairs estimated distance are correlated against the reference network, and adjusted vertex discovery numbers are reported in order to evaluate the performance of the this scheme (pg. 36).

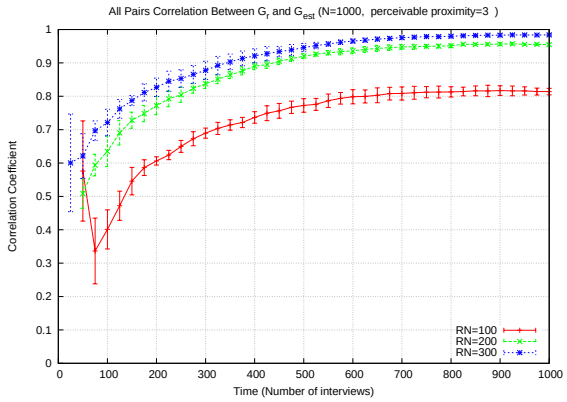
6.1.1 Results for Network of 1000 Nodes. Figure 13 illustrates correlation coefficient values and vertex discovery numbers for a network of 1000 nodes during the course of conducting 1,000 interviews. The figures at the top are obtained from experiments where the perceivable proximity threshold pm=2, while the ones at the bottom are correspond to setup where the perceivable proximity is taken as pm=3.



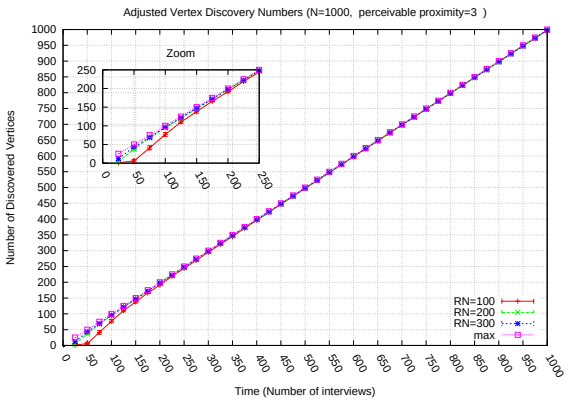
(a) Correlation when pm=2



(b) Vertex discovery numbers when pm=2



(c) Correlation when pm=3



(d) Vertex discovery numbers when pm=2

Figure 13. All pairs correlation and vertex discovery numbers for 1000-node network

The correlation values and vertex discovery numbers are illustrated with error bars which show mean and standard deviation for three recognition numbers rn=100, 200 and 300 with colors red, green and blue, respectively. We shall discuss each of these figures in turn.

Figure 13a shows the correlation coefficient for a network of 1000 nodes when perceivable proximity pm=2. It can be clearly seen that there is an upward trend for all recognition numbers while the slope of increase is larger for the blue and green graphs.

If we analyze the graphs individually, we can see that initial data points for the red graph appear after completing 75 interviews. These data points vary from 0.6 to 0.9 with large errors. From interview number 75 to 150, the red graph drops

sharply from 0.8 to 0.4 with 0.1 standard deviation. By interview number 150, the red graph starts rising and keeps increasing gradually until it reaches a mean correlation coefficient of 0.7 by interview number 1000.

The blue and green graphs show a similar trend. Their initial data points appear after completing just 50 interviews. Between interview number 50 and 100, there is a sharp decline in the correlation value of the blue graph, whereas the green graph experiences a similar drop which it does not overcome until interview number 125. After the blue and green graph "bottom out", they start to show an upward trend, and continue to climb until the end of the experiment, albeit with a decreasing slope.

The interview number at which the three curves reach a correlation coeffient of 0.5, we see that the red graphs reaches this by interview number 150, whereas it takes the blue and green graphs only 50 interviews to reach the same correlation coefficient. When considering a threshold of correlation coefficient of 0.7, we see that the red graph reaches this after 800 interviews while the green and blue achieve after only 400 interviews. By the end of the simulation, the red graph has reached a correlation coefficient of 0.7, while the blue and green graphs have attained a correlation coefficient of 0.9.

We now turn to Figure 13b where we can observe the vertex discovery numbers during the experiments. The purple line which extends from the bottom left to the top right shows the number of subjects that have been interviewed so far, this is also the upper bound of the vertex discovery number. This line is drawn to make it easier to compare the observed values with the upper bound.

We can see that for all recognition numbers, the greatest divergence between the upper bound and the actual vertex discovery numbers occurs during the first 250 interviews. After completing 250 interviews, we see that the graphs essentially converge with the upper bound.

This result may be explained by the fact that when the number of interviews is small, we are forced to use correspondingly small K values. To put it more clearly, in a network of 1,000 nodes, for the interview number is 51, K is equal to 5, whereas for the interview number 251, K is equal to 25 (pg. 29).

In addition, vertex pairs whose pairwise distances are greater than the perceivable proximity threshold are omitted, and as a result they are not included in the estimated network. This feature of the model has an impact on the number of discovered nodes, especially earlier on interview process. For instance, if we let K be 5, none of the selected subjects may be within the range of the perceivable proximity threshold.

Another observation we can make is that all graphs start with high correlation values, but these correlations drop sharply. The largest error bars are also observed at this initial stage when correlations are high. Few vertices are discovered at this early stage, and this results in the falling correlation values and large errors in a few Type-I proxies that are included in the response data. Recall that in the aggregation phase, if a pairwise distance has not been reported , a path between these vertices will be computed, and if a path is found, it will be added as a Type-II proxy within the system. Note that Type-II proxies eventually may cause distortion if there is in reality a shorter path in the reference graph. With these observations in mind, let us now analyze and compare the results of the next set of experiments in which the perceivable proximity threshold is three.

In Figure 13c, we see that the first data points are observed at interview number 25 for the blue, 50 for the red and 50 for the green graphs. The emergence of these data points are earlier than the previously explained experiments.

The red graph shows a sharp decrease between interviews 50 and 75. After showing a dip at interview number 75, the red graph rises sharply until interview number 175. From interview number 175 to 600, the red graph continues to rise

gradually. Between interview number 600 to 1000, red levels out at a correlation value of 0.8.

The Blue and green graphs rise just after their initial data points. They then show a gradual increase until the end of the experiment.

Comparing Figure 13a with Figure 13c, we can see that when we assume larger perceivable proximity thresholds, the model achieves higher correlation coefficient values. For example, while the red graph reaches 0.7 after 300 interviews, in Figure 13a it reaches 0.7 after the 700th interview in Figure 13a. In Figure 13c the blue and green graphs reach above correlation coefficient of 0.5 before completing 100 interview, in Figure 13a, it takes for them 300 interviews to reach the same correlation coefficient.

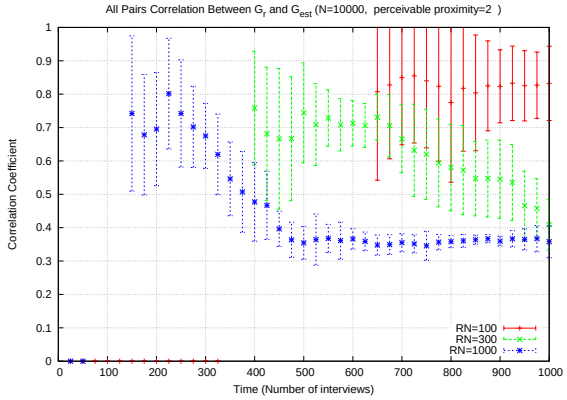
Parallel with this finding, we can see in Figure 13d that the vertex discovery numbers converge to their upper bound earlier. For instance, it takes only half of the number of interviews for the red graph to converge when the model is set to when the perceivable proximity threshold is three.

Taken these findings together, we can conclude that a larger perceivable proximity threshold results in higher vertex discovery numbers and correlation coefficients. A possible explanation of these results may reside in the process with which we add new vertices to the estimated network. Recall that when the distance between vertex pairs is greater than the perceivable proximity threshold, the vertices of this pair are not added in the response data (even if they are selected as objects). It is more likely for two vertices to be at a distance less than or equal to 3 than to be in a distance of 2. Thus, the set of vertices which are reachable at higher proximity threshold settings include vertices which are reachable with lower proximity threshold settings. For this reason, in the experiments with lower perceivable proximity thresholds, vertices are more likely to be omitted from the response data.

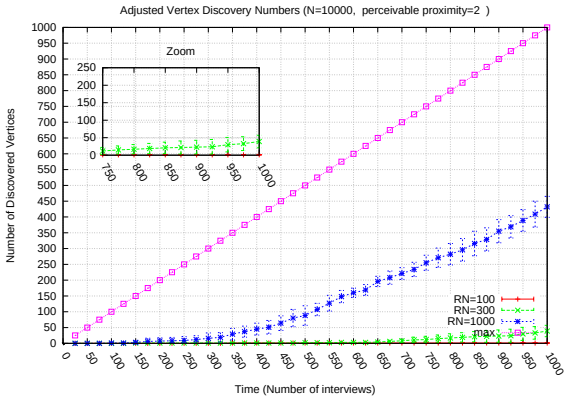
In attempting to explain the rise in the correlation coefficients, we note that

there is an increase in the number of Type-I proxies, and corresponding decrease in the number of Type-II proxies.

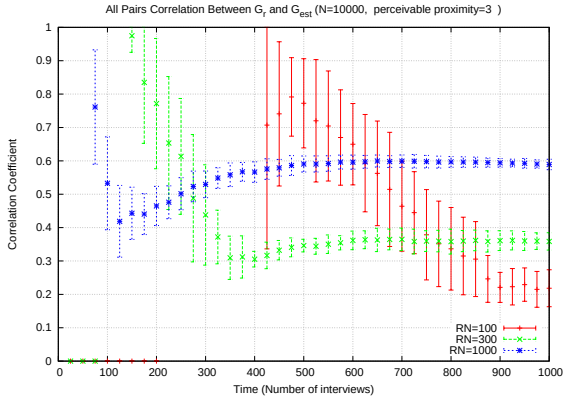
In the next set of experiments, we discuss the analogous findings which are obtained via experiments of networks of 10,000 nodes.



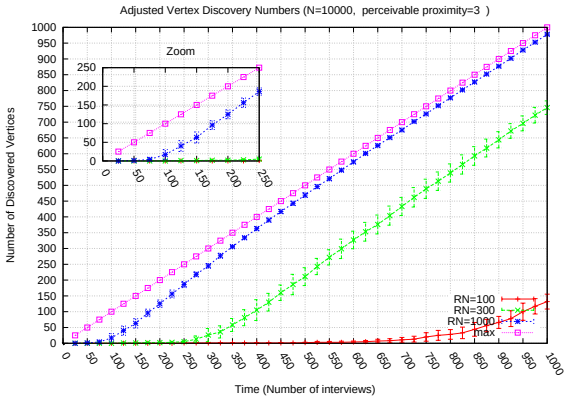
(a) All pairs correlation when pm=2



(b) Vertex discovery when pm=2



(c) All pairs correlation when pm=3



(d) Vertex discovery when pm=3

Figure 14. Model-1 All pairs correlation and vertex discovery numbers for 10,000-node network

6.1.2 Results for Network of 10,000 Nodes. In Figure 14, the correlation values are illustrated for experiments conducted in networks of 10,000 nodes. The recognition number parameter is set to 100,300 and 1000, and the corresponding graphs for these three settings are shown in red, green and blue respectively. Similar to the previous experiment, a total of 1000 interviews is completed by the end of the experiment. Note that 1000 interviews comprise 10% of the total population.

In Figure 14a, the first correlation value for the red graph, is observed after 650 interviews. Then red fluctuates with high deviations until the end of the simulation.

In considering the green graph, we observe that the first correlation coefficient values arise after completing 400 interviews. Then the green graph fluctuates until interview number 675, and continues to drop until the end of the experiment.

When we consider the blue graph, we see that the first data points appear after completing only 150 interviews. From interview number 150 to 450, the blue graph drops gradually, then becomes steady until the end of the simulation.

Figure 14b shows the corresponding vertex discovery rates. We can see that the vertex discovery numbers are very low. For the blue graph, where the recognition number is 1000, the number of discovered vertices rises with a steady pace between interview number 300 and 1000, and finally reaches approximately 400 vertices by the end of the simulation.

Figure 14b reveals that we discover very few vertices when the recognition number is 100 or 300. When the recognition number is 1000, the correlation value is observed to be 0.4 after we have completed interviews of 10% of the total population.

In the experiments when perceivable proximity threshold is assumed to be three, we can see that initial data points appear earlier, at interview number 75, 150 and 425 for blue, green and red graphs respectively. The red graph shows a downward trend with large errors during the experiment and indicates correlation value of 0.2 by the end. The green graph drops until interview number 350, reaching correlation value of 0.3 ± 0.07. Between interview number 350 to 1000, the green graph exhibits correlation values around 0.3. Blue starts rising after interview number 125 and gradually increases until 500. Between interview number 500 and

1000, the blue graph shows a steady pattern which levels off at correlation value of 0.6.

In the next part of this Section, we will compare the 1000 and 10,000 node experiments.

6.1.3 1000 vs 10,000-node Networks. Comparing the experiments of 1000 and 10,000-node graphs, we can observe a significant decrease in vertex discovery numbers and correlation values for larger networks.

A possible explanation for this might be the decrease in the recognition rate in larger networks. Recall that the recognition ratio determines the number of objects to be selected in an interview, and is calculated as

$$\text{Recognition ratio } = \frac{\text{Recognition Number}}{\text{Network Size}}.$$

Given this formula, we can say that any increase in network size causes a linear decrease in the recognition ratio. For instance, let the interview number be t= 200 , the recognition number is 100. In a network of 1000 nodes, the recognition ratio will be 0.10 whereas a 10,000-node network the recognition ratio will be only 0.01. If we calculate K for these two scenarios, the K will be 20 for the 1000-node network, but the K will be 2 for 10,000-node graphs. In Table 8, K values for different number of interviews are shown for each of the two network sizes.

We see from Table 8 that at the same interview number, in larger networks, the number of discovered vertices is lower. One should consider the fact that the K value is an important factor in determining the efficiency of vertex discovery during the data collection process. Another reason for the decrease in efficiency with larger networks may be the subject selection scheme (random sampling without replacement). In the implementation of these experiment, subjects are selected randomly from the reference network. Given that the objects are selected from

t \ RN	100	200	300
100	10	20	30
200	20	40	60
300	30	60	90
400	40	80	120
500	50	100	150
600	60	120	180
700	70	140	210
800	80	160	240
900	90	180	270
1000	100	200	300

a)Network of 1000 nodes

t \ RN	100	300	1000
100	1	3	10
200	2	6	20
300	3	9	30
400	4	12	40
500	5	15	50
600	6	18	60
700	7	21	70
800	8	24	80
900	9	27	90
1000	10	30	100

b)Network of 10,000 nodes

Table 8

Number of objects to be selected at each 100 interviews according to different recognition numbers

among the prior subjects, we can infer that randomly selected objects from a larger network will more likely to be further away from each other than that vertices selected from smaller networks. Accordingly, we would expect that the vertex pairs are more likely to be omitted in the estimated graph because their pairwise distances are greater than the perceivable proximity threshold.

We also see higher error rates in the correlation value when considering the 10,000-node network. The reason for these higher errors bars may be that fewer pairs of subjects are compared during the experiments (This is particularly true when the recognition number is small). In larger networks, Type-II proxies are more likely to produce greater distortion if the Type-II proxy is not coincident with shortest path.

To sum up, Model-1 performs better in 1000-node networks than in 10,000-node networks. In a 10,000-node network when perceivable proximity threshold is 2, vertex discovery numbers are very low, especially for recognition number 100 and 300. When the perceivable proximity threshold is taken to be 3, at recognition number 1000, the model can reach correlation values above 0.5.

pm	RN	$\rho \geq 0.5$	$\rho \geq 0.7$
2	100	350	850
2	200	300	425
2	300	175	375
3	100	150	325
3	200	50	150
3	300	25	100

a)Network of 1000 nodes

pm	RN	$\rho \geq 0.5$	$\rho \geq 0.7$
2	100	few nodes	few nodes
2	300	few nodes	few nodes
2	1000	-	-
3	100	-	-
3	300	-	-
3	1000	500	-

b)Network of 10,000 nodes

Table 9

The table shows the minimum number of interviews necessary to reach a correlation coefficient greater than 0.5 and 0.7 (Model-1)

6.1.4 Summary of Findings. In Table 9, it can be seen that Model-1 performs better in 1000-node networks compared to 10,000-node networks. In a 1000-node network, Model-1 can reach above a correlation value of 0.5 with all recognition numbers by interviewing at most 350 person in the population. The number of interviews that are necessary decreases to 150 when the perceivable proximity threshold is taken two. By examining the figure we may conclude that:

- Recognition ratio which determines the number of objects shown is positively correlated with the model's performance,
- Perceivable proximity threshold has a direct effect on the performance,
- Network size is significantly and negatively correlated to the performance.
- Random sampling method is more likely to select objects which are further away from each other, which results in lower vertex discovery rates and correlation coefficients
- Model-1 is inefficient for 10,000-node networks, if the recognition ratio is less than 10% of the population size.

6.2 Model-2

In Model-1, subject selection was performed with random sampling in which all vertices were given the same chance for being selected. While random sampling

is a well accepted method, this scheme may result in choosing subjects which are further from each other, and as discussed above, this results in omitting vertices from the estimated graph.

Considering this consequence of random sampling, in Model-2 we used Respondent Driven Sampling instead. Throughout the experiments, RDS is implemented with following parameters.

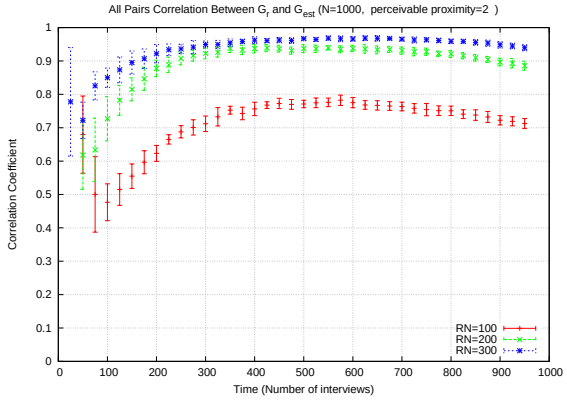
- Number of initial seeds for RDS s= 10
- RDS coupons c=3

The tests for Model-2 make use of the reference graphs which are identical to the ones used in testing Model-1. In the discussion below, we examine the results of the same experiments that are used to test Model-1, but here the focus is describing the impact of using the RDS scheme. At the end of this part, we will also provide a comparison of Model-1 and Model-2.

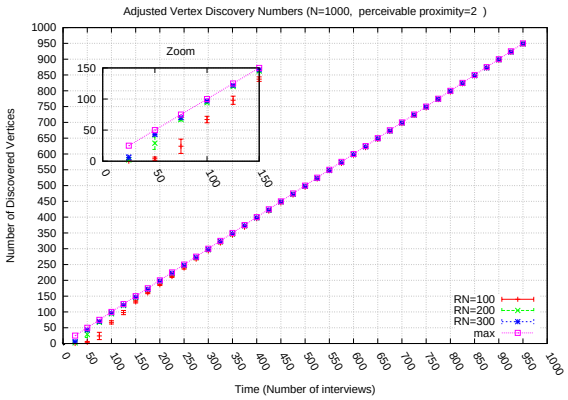
6.2.1 Results for Network of 1000 Nodes. Figure 15 illustrates all pairs correlation values between estimated network and reference network, and as well as, the vertex discovery numbers during the course of 1000 interviews.

In Figure 15b, we observe that the fist data point of the red graph appears at interview number 50. The red graph exhibits a dip, at interview number 100, then rises gradually until interview number 425 where it achieves a correlation coefficient of 0.75. The red graph continues to remain above correlation coefficient 0.7 until the end of the simulation, although it experiences a slight decline after interview number 500. For the green and blue graphs, we observe a gradual increase between interview numbers 100 and 200. After interview number 200, both graphs level out at a correlation coefficient in excess of 0.9.

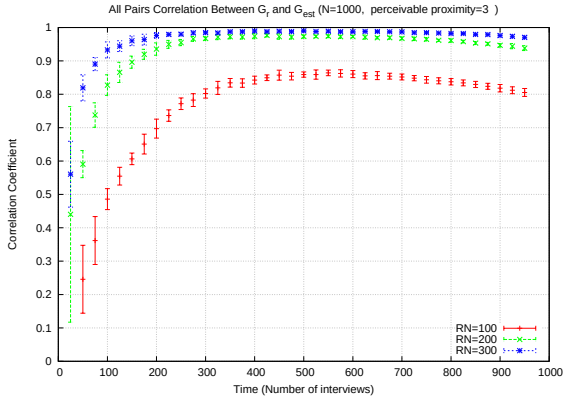
We now turn to vertex discovery rates. In Figure 15b, we can see that the graphs converge with the upper bound vertex numbers rapidly. The only exception



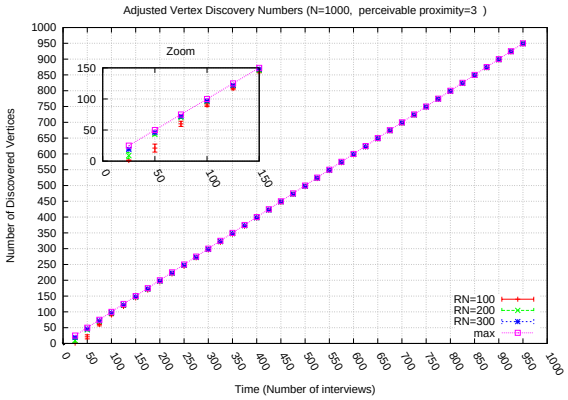
(a) All pairs correlation when pm=2



(b) All pairs correlation when pm=2



(c) All pairs correlation when pm=3



(d) All pairs correlation when pm=3

Figure 15. Model-2 All pairs correlation and vertex discovery numbers for 10,000-node network

may be the performance of the red graph during the first 150 interviews. We see that at interview number 50, the red graph has only discovered approximately 10 vertices. The number of discovered vertices becomes nearly 60 when the red graph has completed 100 interviews. By interview number 150, the red graph has essentially converged with the upper bound number.

When we analyze the number of interviews that these three graphs require in order to reach a particular correlation values, we can see that after 300 interviews, the red graph reaches above a correlation coefficient of 0.7. At interview number 300, red's vertex discovery number achieves the upper bound value. For the blue and green, the discovered vertices converge with the upper bound after only 100 interviews.

We now turn to Figures 15c and 15d, where we assume perceivable proximity threshold is three. It can be seen that all data points approximately become 0.1 point higher than in this scenario when the perceivable proximity threshold is two. After the 225^{th} interview, the red graph reaches a correlation coefficient 0.7. It ends with a correlation coefficient in excess of 0.8 at interview number 1000. The blue and green graphs show correlation coefficient values of 0.9 only after completion of 175 interviews.

Note that in Figure 15, the last data points are observed at the 950^{th} interview, because RDS has already selected all possible subjects, and the vertex set under consideration is exhausted. In the next part of this section, we will discuss the results of the corresponding experiments which are conducted on networks of 10,000 nodes.

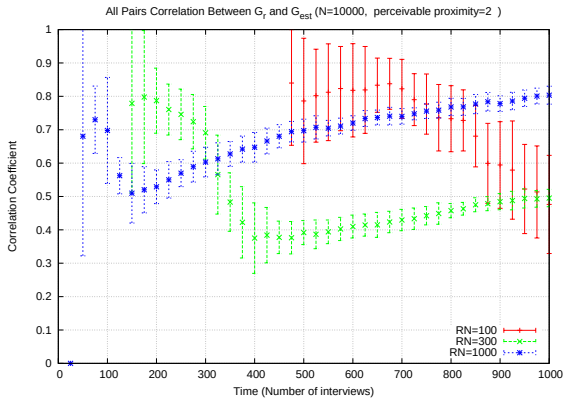
6.2.2 Results for Network of 10,000 Nodes. Figure 16 shows correlation coefficient and vertex discovery numbers when Model-2 is applied to networks of 10,000 nodes.

One of the noteworthy features observed in Figure 16b is the size of error bars. In addition, the error bars are smaller when the recognition number is larger. Another observation is the trend: all graphs experience a fall, and then rise continuously as the interviews proceed.

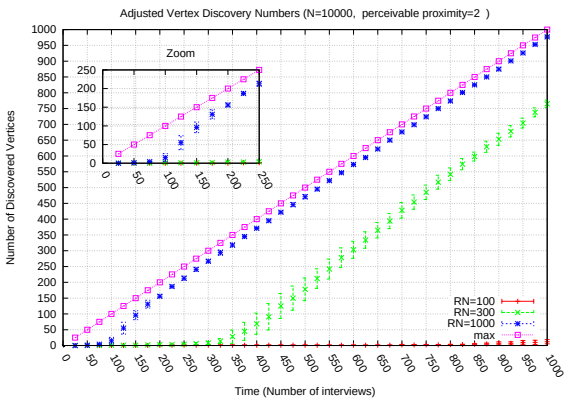
When we consider the vertex discovery rate, we see that the red graph only discovers a few vertices until it has completed 1000 interviews, and the correlation coefficient exhibits a decreasing trend throughout.

When we consider the green graph, we see that the correlation coefficient value is only 0.4 at interview number 400. Starting from this interview onwards, the green graph rises steadily, and at the end of 1000 interviews it reaches a correlation coefficient of 0.5.

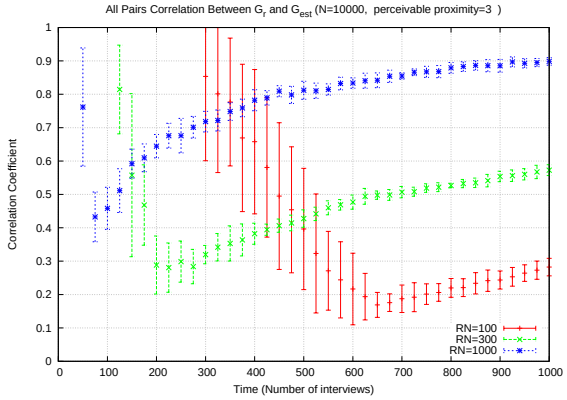
As expected, we observe the best performance when the recognition number



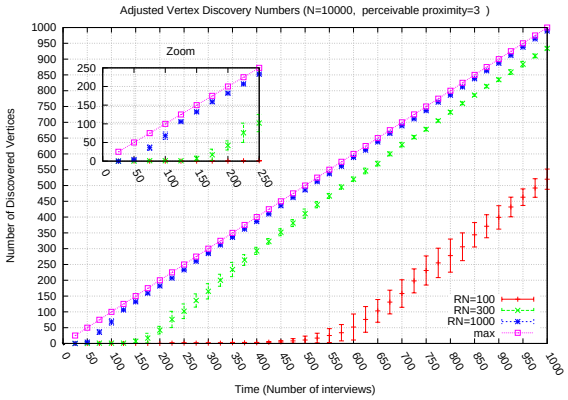
(a) All pairs correlation when pm=2



(b) Vertex discovery numbers when pm=2



(c) All pairs correlation when pm=3



(d) Vertex discovery numbers when pm=3

Figure 16. Model-2: All pairs correlation and vertex discovery numbers for 10,000-node network

is the highest. The blue graph fluctuates in the first 150 interviews, and then begins to rise steadily after interview number 150 at which point the correlation coefficient is 0.5. This steady rise in correlation coefficient continues as interviews proceed finally reaching a value of 0.9 by interview number 1000. Next we consider how these experimental results change when the perceivable proximity threshold is increased to three.

Figure 16c and 16d, illustrate the performance of accuracy of the estimated network when perceivable proximity threshold is taken to be three.

The figures show that higher correlation values and vertex discovery numbers are achieved for all settings of recognition number parameter. After completing 1000 interviews, the red graph reaches a correlation coefficient of 0.3, and has

discovered 500 vertices. In comparison, the green graph has achieved a correlation coefficient in excess of 0.5, and has discovered 950 vertices. Finally the blue graph has achieved a correlation coefficient of 0.9, and has discovered 1000 nodes.

pm	RN	$\rho \geq 0.5$	$\rho \geq 0.7$
2	100	150	300
2	200	50	125
2	300	25	75
3	100	125	225
3	200	50	75
3	300	25	50

a)Network of 1000 nodes

pm	RN	$\rho \geq 0.5$	$\rho \geq 0.7$
2	100	-	-
2	300	1000	-
2	1000	200	575
3	100	-	-
3	300	675	-
3	1000	175	300

b)Network of 10,000 nodes

Table 10

The table shows the minimum number of interviews necessary to reach a correlation coefficient greater than 0.5 and 0.7 (Model-2)

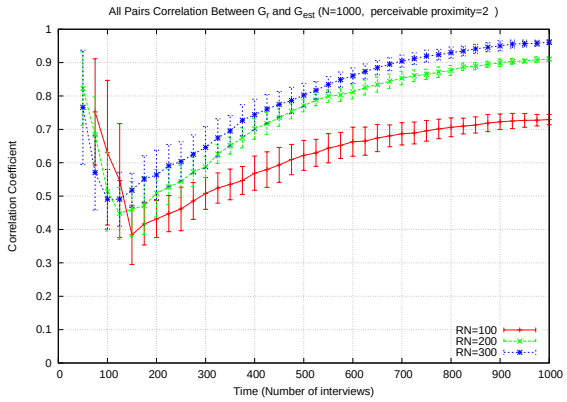
6.2.3 Comparing Model-2 vs Model-1. In Table 10, we can see that in 1000-node networks, the correlation coefficient rapidly reaches above 0.7 even when the perceivable proximity threshold is assumed to be 2. For instance, with the smallest recognition number, it takes 300 interviews for Model-2 to reach above correlation coefficient of 0.7. On the other hand, the blue and green graphs indicate higher correlation values in excess of 0.7 at just 150 interviews.

For 10,000-node networks, when the perceivable proximity threshold is 2, the correlation coefficient values are very low when the recognition number is 100. However, when the recognition number is 300 or 1000, Model-2 attains correlation coefficients above 0.5 with vertex discovery numbers of 200 and 1000, respectively.

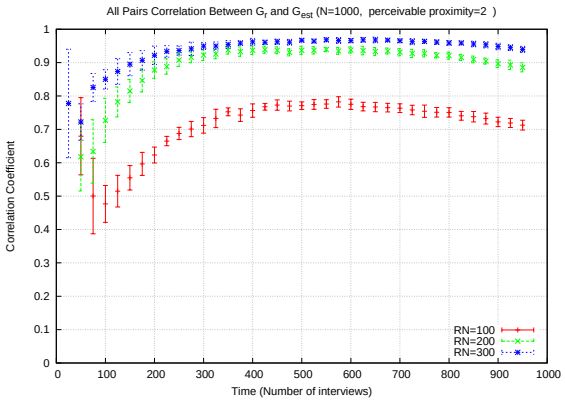
By the end of the interview process, the green graph (where the respondents recognize 3% of the population) the correlation coefficient reaches 0.5. When the recognition ratio is increased to 10% (as is the case in the blue graph) the correlation coefficient achieved is much higher, 0.8. Considering the trend across graphs in these experiments, we believe the correlation coefficient will continue to rise as we interview more subjects.

In Figure 17, the results of Model-1 and Model-2 are provided. The figures at

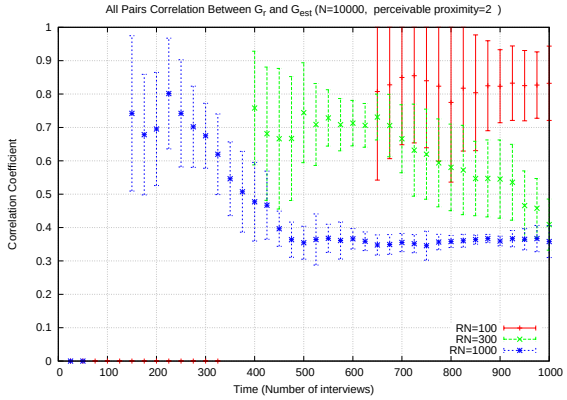
the top show results of 1000-node networks, and the figures in the second row illustrate the 10,000-node-networks experiments. For simplicity, we do not show all the results for Model-1 and Model-2, and only compare those where the perceivable proximity threshold is taken to be 2.



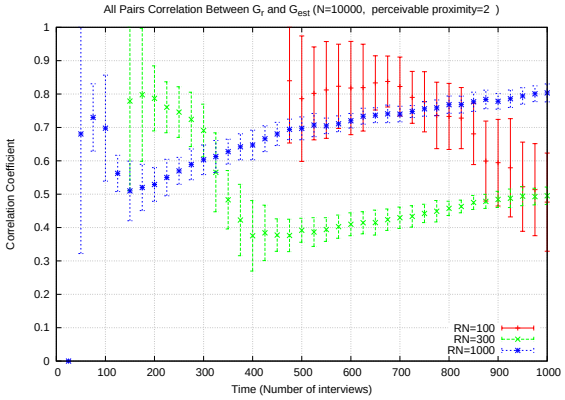
(a) Model-1 All pairs correlation for 1000-node network



(b) Model-2 All pairs correlation for 1000-node network



(c) Model-1 All pairs correlation for 10,000-node network



(d) Model-2 All pairs correlation for 1000-node network

Figure 17. Comparison of Model-1 and Model-2 when perceivable proximity threshold is assumed to be 2

By comparing the figures at the top, we see that the final correlation coefficient values of both models are commensurate. However, in the Model-2 experiments, the graphs achieve their maximum correlation values more rapidly as the interview process proceeds. For instance in Model-1, the red graph reaches correlation coefficients of 0.5 and 0.7 at interview number 300 and 700, respectively. In Model-2, we reach these correlation coefficient values at interview numbers 125

and 300, respectively. In other words, Model-2 achieves the same correlation coefficient values after fewer interviews than Model-1 does. This suggests that in 1000-node networks, Model-2 is significantly more efficient than Model-1.

In 10,000-node network experiments, we observe that when recognition ratio is 1% (as is the case in the red graph), none of the models can achieve a high correlation. On the other hand, when the recognition number is taken to be 300 or 1000, Model-2 reveals higher correlation values (0.5 and 0.8, respectively) after completing 1000 interviews.

To summarize, we conclude that Respondent Driven Sampling performs significantly better than the random sampling. The main reason for this improvement is that RDS favors the selection of subjects which are local to one another in the network. Since objects are selected from previously interviewed subjects, this bias in RDS ensures that a greater percentage of objects will lie within the perceivable proximity threshold, and thus be included in the estimated graph.

6.2.4 Summary of Findings.

- Respondent Driven Sampling enhances the performance of the model in all experiments,

- In 1000-node networks, the models can reach correlation coefficients of 0.5 and 0.7 after recruiting 150 and 300 respondents, respectively. When we have recognition numbers of 200 and 300, Model-2 reaches a correlation coefficient in excess of 0.7 interviewing only 125 respondents (page 58).

- In 10,000 node networks, after interviewing 10% of the entire population, Model-2 attains a correlation coefficient in excess of 0.5 when the recognition number is 300. When the recognition number is increased to 1000, this correlation coefficient increases to 0.8.

6.3 Model 3

In previous results, we saw that RDS performs much better than random sampling, especially in large networks. For this reason, in Model-3, we retain RDS as the subject selection scheme. However, in Model-3, the object selection is performed by sampling within Ego Network First(ENF) instead of random sampling. In the ENF scheme, we seek to minimize the omitted reports by favoring the selection of objects that are closer to one another.

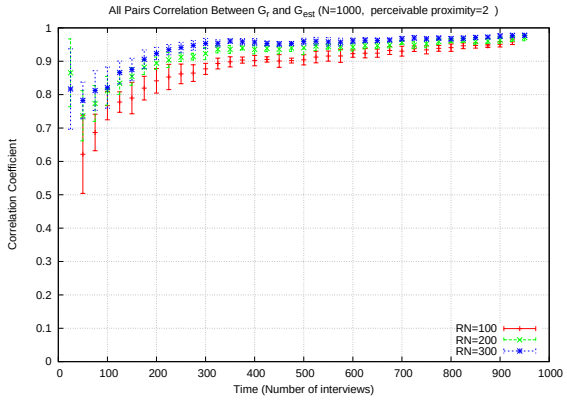
In order to achieve this goal, the respondents are asked to report the distance between themselves and previous recruits. Then the system selects a set of objects which are close to the subject, and asks the subject to estimate the perceived proximity between these objects. The reader may refer the description in pg. 31 for a more detailed explanation of the ENF.

Note that in this process, we retain our assumption that subjects can only report the distance between a pair of objects if the distance is less than or equal to the perceivable proximity threshold. A respondent who is asked to estimate the distance between his/herself and the object will only be able to do so if this distance less than or equal to the perceivable proximity threshold.

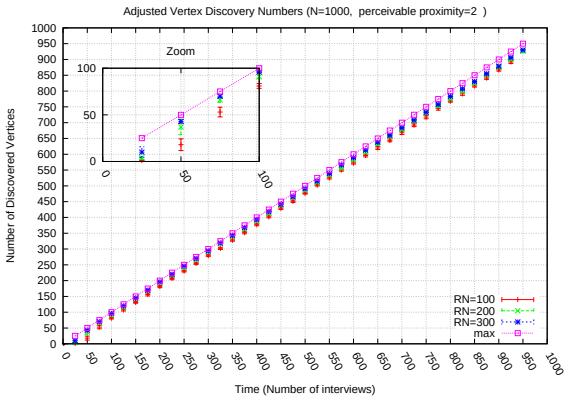
With respect to privacy issues, Model-3 presents some challenges, because respondents are being asked to reveal their own relationships. However, this information is only used for object selection, and is not incorporated incorporated into the estimated network structure itself. Thus, subjects may be reassured by the promise that the information they provide about their own relationships is not being retained in any way by the system.

As in previous experiments, the tests for Model-3 are based on the same reference graphs that are used to test the previous two models. After presenting the results of Model-3 tests, we will make a comparison between Model-2 and Model-3.

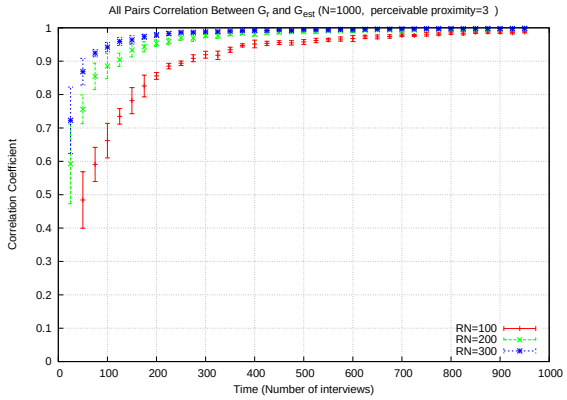
6.3.1 Results for Network of 1000 Nodes. In Figure 18a and Figure 18b, the results are for the situation where the perceivable proximity threshold is 2. In Figure 18a, all graphs exhibit a correlation coefficient in excess of 0.5 even at the very beginning of the interviewing process. Just after the first data points appear, the graphs start to rise gradually, and reach values above 0.7 after only the first 100 interviews. They then continue to climb until interview number 400, and from this point on, the graphs always remain above 0.9 showing minor fluctuations.



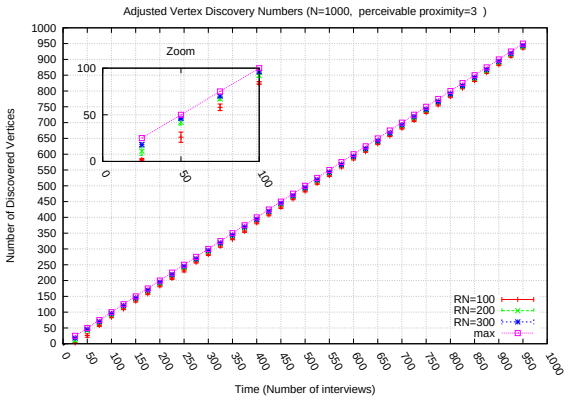
(a) All pairs correlation when pm=2



(b) Vertex discovery numbers when pm=2



(c) All pairs correlation when pm=3



(d) Vertex discovery numbers when pm=3

Figure 18. All pairs correlation and vertex discovery numbers for 1000-node network

When we look at Figure 18b, we see that the vertex discovery numbers converge to the upper bound vertex numbers after only 100 interviews. We can also observe that the red graph shows large errors prior to its convergence to the upper bound.

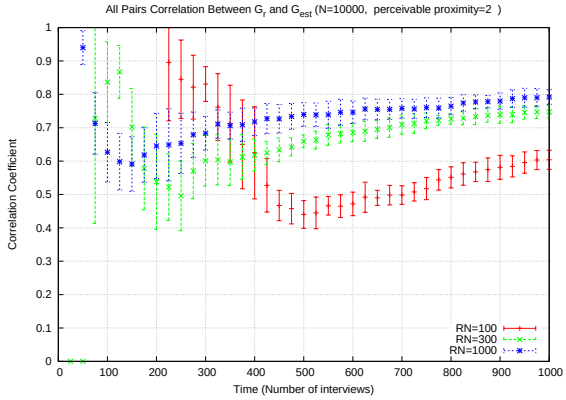
When we compare the correlation coefficient values achieved under different perceivable proximity threshold settings, we can observe that increasing the perceivable proximity threshold enhances the rate at which the maximum correlation coefficient value is reached as the interview process unfolds. For example, in Figure 18c when the perceivable proximity threshold is three, the blue, green and red graphs reach 0.9 at interview number 100, 125 and 275, respectively. In comparison when the perceivable proximity threshold is set to 2, these same correlation coefficients requires 175, 200 and 375 interviews, respectively to reach 0.9.

If we compare vertex discovery numbers, we can see that the impact of the perceivable proximity threshold is quite small compared to its impact in the previous two models. Comparing the red graph in Figure 18b and Figure 18d, we see that both configurations reveal approximately the same vertex discovery numbers. In the next part of this Section, we present the results of analogous experiments conducted on a network of 10,000 nodes.

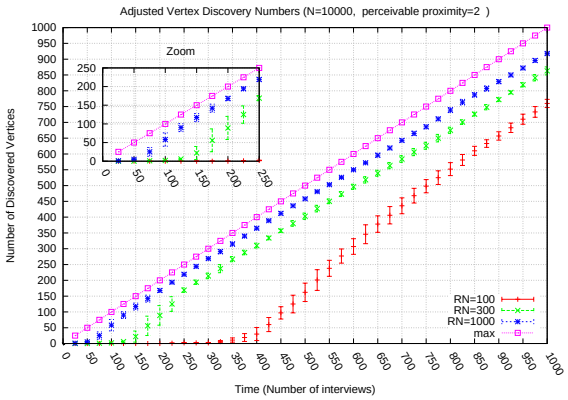
6.3.2 Results for Network of 10,000 Nodes. Figure 19 shows the experimental data from a 10,000-node network experiment in which Model-3 is applied.

Similar to previous models, in Figure 19a, the correlation graphs show a trend that begins with large fluctuations. These fluctuations continue until the graphs bottom out. After the graphs experience an initial dip, they theny rise steadily throughout the interview process. The red graph reaches a correlation coefficient of 0.6, while the green and blue graphs reach correlation coefficient in excess of 0.7 upon completing 1000 interviews. Note that Model-3 provides the highest correlation coefficient value for the red graph (which corresponds to a situation in which respondents can recognize only 1% of the total population).

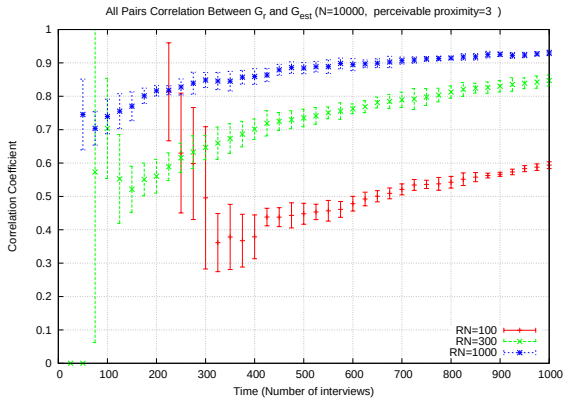
With respect to vertex discovery numbers,the red, green and blue graphs discover 750, 850 and 950 vertices, respectively in the course of 1000 interviews.



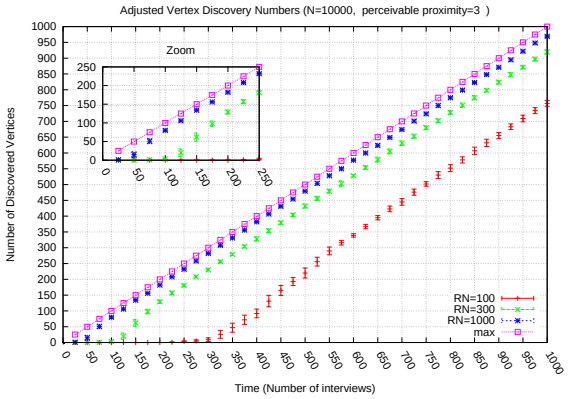
(a) All pairs correlation when pm=2



(b) Vertex discovery numbers when pm=2



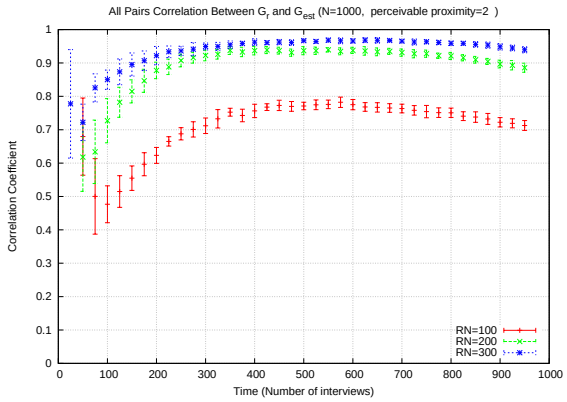
(c) All pairs correlation when perceivable pm=3



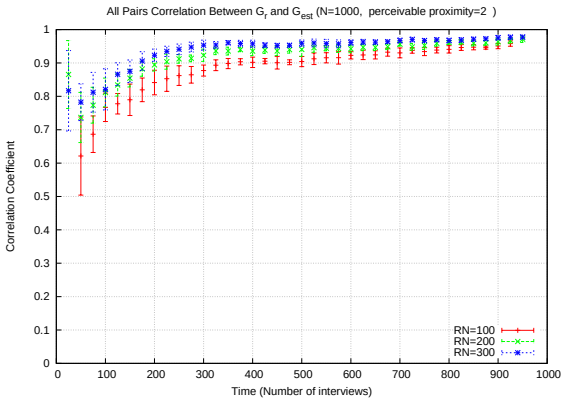
(d) Vertex discovery numbers when pm=3

Figure 19. Model 3 All pairs correlation and vertex discovery numbers for a 10,000-node network

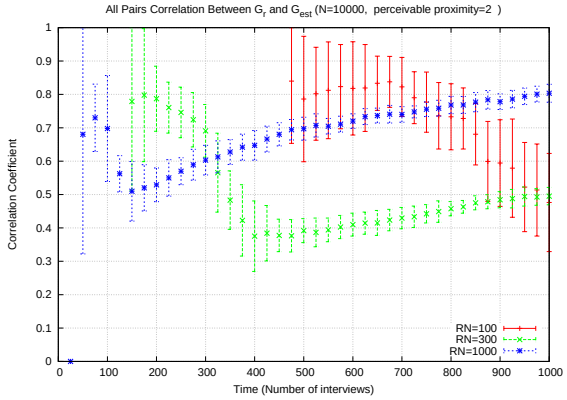
Comparing the performance of Model-3 with respect to changes in the perceivable proximity threshold, we see similar trends in the correlation coefficient graphs. When the perceivable proximity threshold is increased to 3 the correlation values for all graphs are approximately 0.1 higher than when the perceivable proximity threshold is 2. Another marked difference is that the error bars become significantly smaller when the perceivable proximity threshold is increased.



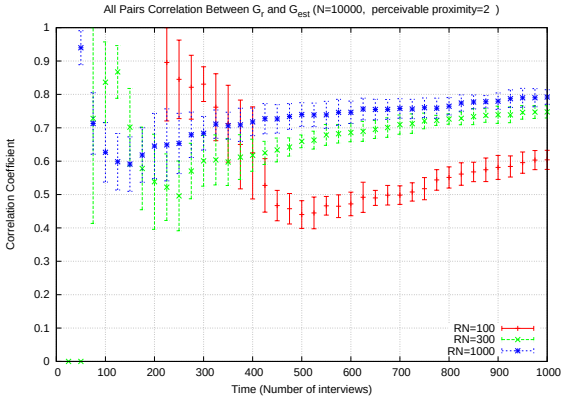
(a) Model-2 All pairs correlation for 1000-node network



(b) Model-3 All pairs correlation for 1000-node network



(c) Model-2 All pairs correlation for 10,000-node network



(d) Model-3 All pairs correlation for 1000-node network

Figure 20. Comparison of Model-2 and Model-3 when perceivable proximity threshold is assumed to be 2

6.3.3 Comparing Model-3 vs. Model-2. We will now present a comparison between Model-2 and Model-3 for both 1000 and 10,000-node networks. Throughout this comparison, we will assume the perceivable proximity threshold is taken as 2; similar results hold when the perceivable proximity threshold is taken to be 3.

When we compare Figure 20a to 20b, the blue and the green graphs show a similar trend. However, we can see that there is a remarkable increase in the correlation coefficient values of the red graph in Model-3. We see that the red graph reaches a correlation coefficient of 0.7 by interview number 100, and continues to rise to 0.8 for the next 100 interviews. In Model-2 the red graph reaches a

correlation coefficient of 0.7 but it requires 300 interviews to do so.

When we turn to the experiments in networks of 10,000 nodes, we see that Model-3 performs significantly better than Model-2, especially for red and green graphs. In Model-2, the red graph at the end of the interview process still exhibits large error bars, and shows a downward trend. In Model-3, in contrast we see that the red graph reaches a correlation coefficient of 0.6 by the end of the interview process with an upward trend.

When we look at the green graph, we also see that Model-3 outperforms in terms of correlation values. The green reaches above 0.7 by the end of the experiment, whereas in Model-2 it reaches only 0.5.

The blue graph, on the other hand, exhibits behavior that is comparable in both Model-2 and Model-3, reaching a correlation coefficient of 0.8.

To conclude, when the recognition numbers are low, Model-3 significantly outperforms Model-2 in terms of correlation coefficient values and vertex discovery numbers. This advantage decreases when the recognition number is assumed to be larger.

pm	RN	$\rho \geq 0.5$	$\rho \geq 0.7$
2	100	50	100
2	200	25	25
2	300	25	75
3	100	75	125
3	200	25	50
3	300	25	50

a)Network of 1000 nodes

pm	RN	$\rho \geq 0.5$	$\rho \geq 0.7$
2	100	700	-
2	300	275	750
2	1000	75	425
3	100	650	-
3	300	175	425
3	1000	50	125

b)Network of 10,000 nodes

Table 11
The table shows the minimum number of interviews necessary to reach a correlation coefficient greater than 0.5 and 0.7 (Model-3)

6.3.4 Summary of Findings.

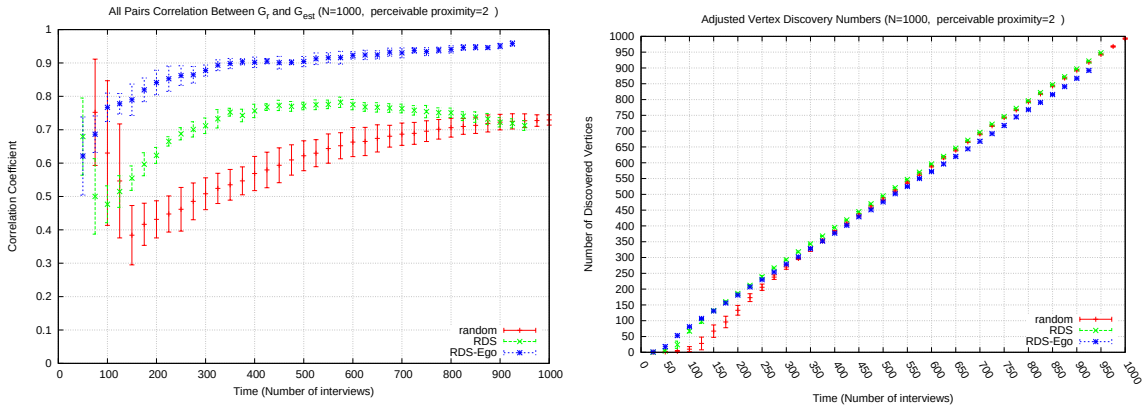
- In 1000-node networks, it takes at most 100 interviews for Model-3 to reach a correlation coefficient in excess of 0.7.
- In 10,000-node networks, where the recognition number is 100, after interviewing 10% of the entire population, Model-3 can reach a correlation coefficient of 0.5. When the recognition number is increased to 300 or 1000, the model achieves a correlation coefficient in excess of 0.7.

7 Discussion

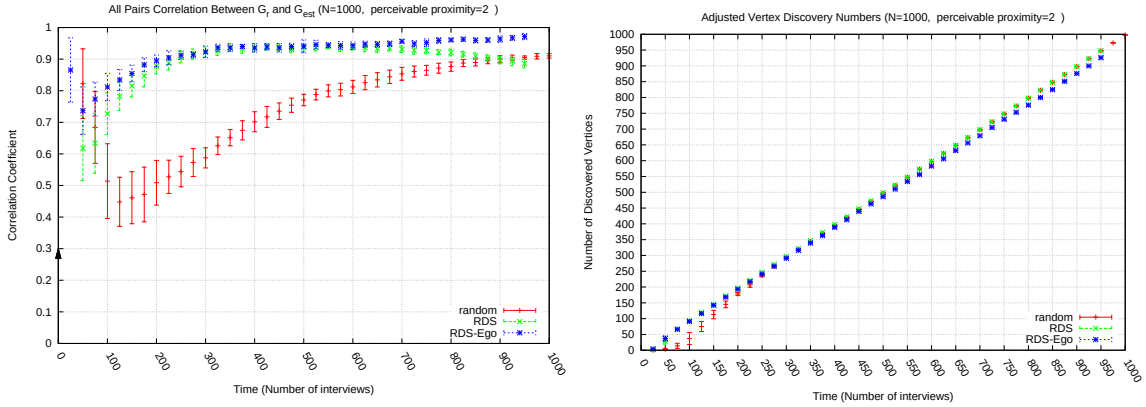
In this section, we will compare the results of each model in turn with respect to correlation values, vertex discovery rates, time efficiency and anonymity assurance. In contrast with the previous sections where we examined each model separately, in the figures that follow, we will illustrate all three models together in order to make evident the differences between them.

In Figure 21, the results are presented for networks of 1000 nodes. The red, green and blue graphs represent random (Model-1), RDS (Model-2), and RDS-Ego (Model-3), respectively.

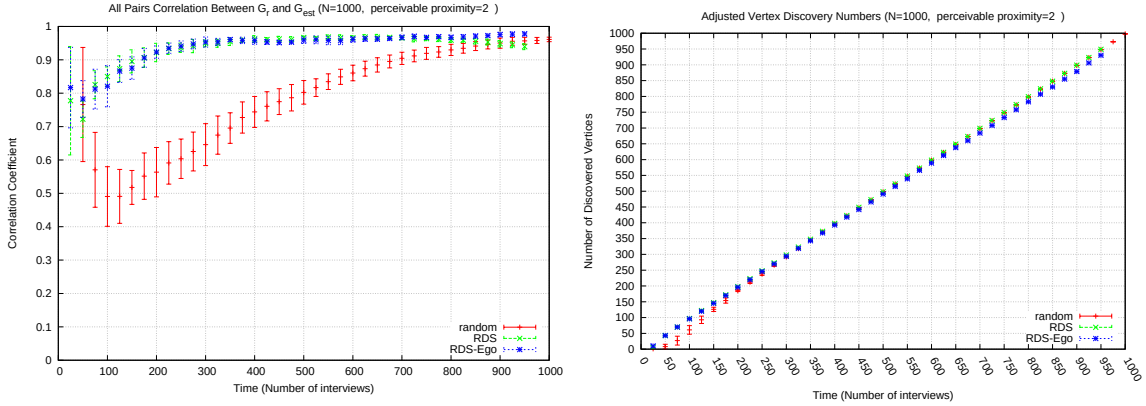
When we analyze the correlation values for different recognition number settings, we can see in Figure 21a that when the recognition number is 100, Model-3



(a) Recognition number=100



(b) Recognition number=200



(c) Recognition number=300

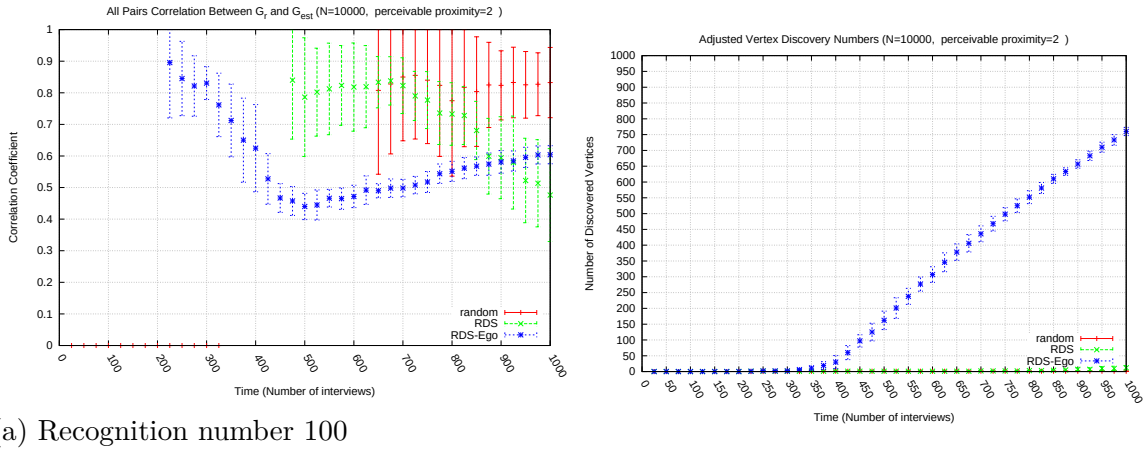
Figure 21. Comparison of results for network of 1000 nodes (Perceivable proximity =2)

performs significantly better than Model-1 and Model-2. In Figure 21b and Figure 21c, when the recognition numbers is increase to 200 or 300, we can see that Model-2 and Model-3 perform almost identically in terms of correlation values. We also see that the increase in recognition ratio decreases the performance gap between

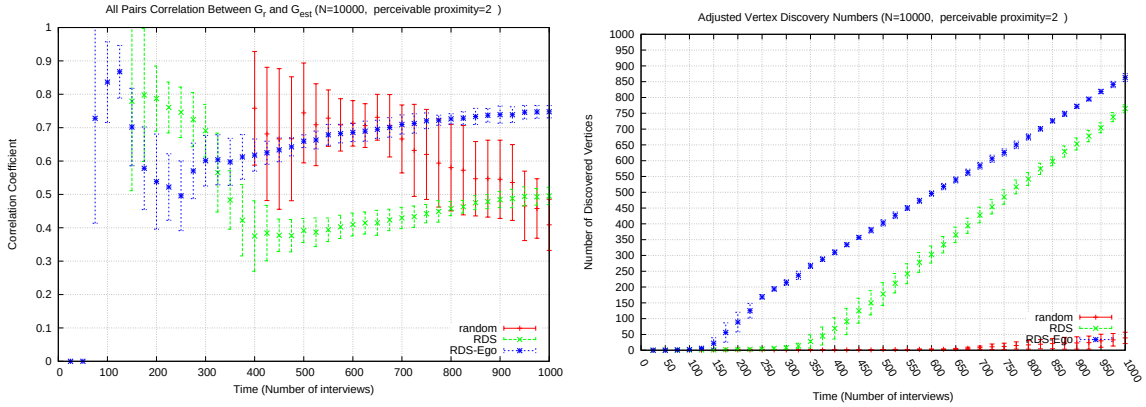
Model-2 and Model-3. The increase in the recognition ratio raises the size of the object set, which ultimately increases the number of Type-I proxies. From this fact, we can conclude that when higher rates of recognition are expected, either Model-2 or Model-3 should be chosen. This design choice becomes more important especially when network privacy and anonymity concerns might make Model-3 unattractive.

When we consider the vertex discovery numbers, we see that Model-2 and Model-3 exhibit comparable performance while Model-1 exhibits much lower performance, especially in the initial stages of the interview process.

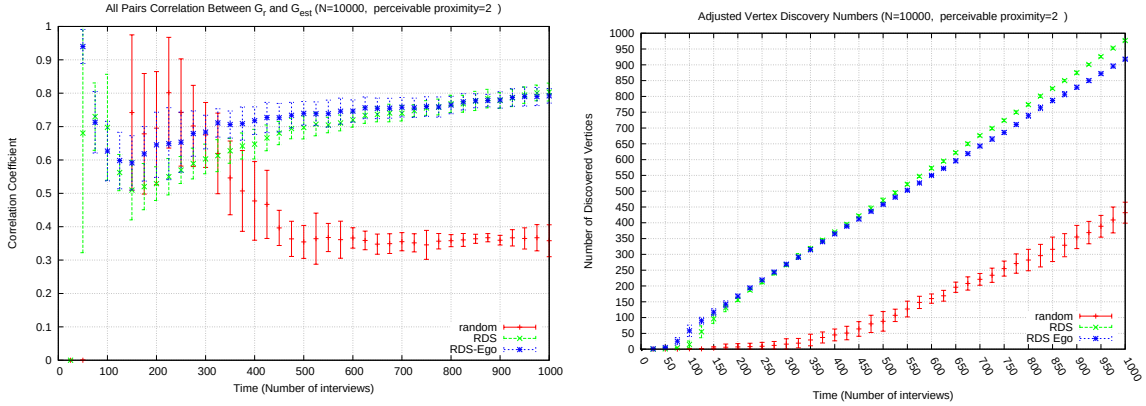
Analyzing the overall results from the 1000-node networks, we conclude that the performance is most greatly impacted by the subject selection scheme and the recognition number. The RDS scheme which is implemented in both Model-2 and Model-3 yields higher vertex discovery numbers and higher correlation coefficient values. As noted before, RDS enables the sample to grow in such a way that recruits are closer to previous respondents. This feature increases the chances of RDS based models to select random objects that are within the perceivable proximity threshold. Thus, RDS based schemes result in greater numbers of Type-I proxies, and they can achieve higher vertex discovery numbers and correlation values.



(a) Recognition number 100



(b) Recognition number 300



(c) Recognition number=1000

Figure 22. Comparison of results for network of 10,000 nodes(Perceivable proximity =2)

If we now turn to network of 10,000 nodes, we observe that when the expected recognition number is 1% of the entire population, Model-3 is the only one which can exhibit acceptable vertex discovery numbers. Comparing Model-2 to Model-3, wherein the only difference is the object selection scheme, we can see that

preferring objects that are close to the subject (as is done in Model-3) results in selecting vertex pairs that are close to one another which increases the likelihood to choose objects within the perceivable proximity threshold. This leads to Model-3's superior performance and the discovery of 750 vertices with a correlation coefficient of 0.6 after completing 1000 interviews. When we look at Figure 22c, where the recognition number is 300, we observe that Model-2 discovers 750 vertices with the correlation value of 0.5 where Model-3 discovers 850 vertices and reaches a correlation coefficient value of 0.7.

In the last set of experiments, when the recognition rate is assumed to be 10%, we see that Model-2 and Model-3 exhibit commensurate performance, and the vertex discovery numbers are above 900 and correlation coefficient values are around 0.8. Like our experiments in 1000-node network, we see that the increasing the recognition ratio enhances Model-2's performance significantly (and more than the improvement experienced in other models.)

Taking together all the experiments discussed so far, we can summarize our observations as follows:

- RDS enables the models to sample individuals who are close to each other. This ability of RDS enhances the performance of the selector to choose objects within the perceivable proximity threshold.
- In larger networks, the probability for two vertices to be within a geodesic distance less than or equal to the perceivable proximity threshold is lower than the same probability in smaller networks.
- During object selection, selecting objects that are close to the respondents results in better performance. This strategy, however, requires respondents to report some information about their ego-networks.

When evaluating the performance of the three models we should take into

consideration time efficiency. This is dependent on the number of objects, K, which are shown in interviews. Recall that the respondents are asked to report $\binom{K}{2}$ pairwise distances. This implies that the time necessary to conduct the interview and the number of recognized objects follow a quadratic relation. Considering the fact that the roster method becomes unmanageable after 30-50 names (Butts, 2008), organizing object pairs in such a way that respondents can quickly report the distance between pairs is a challenging design objective.

With regards to privacy and anonymity, in all three of the models, respondents are asked to report perceived proximity among alters who have been already enumerated in the sample. In other words, respondents are asked to report social proximity among the subjects who are already recorded in the system. In terms of privacy concerns, Model-1 and Model-2 are superior as no information is asked of subjects concerning their own social ties. In Model-3, the subjects are asked to report their social distances from the people who have been interviewed so far. Although this information is not stored and is only used for selecting objects with a bias towards the individuals that are closer to the interviewee within the social network, we note that it might still raise some privacy concerns.

8 Limitations and Future Research

There are a number of potential limitations to this work which we hope to address in time.

We assume that respondents will be able to report the perceived social distance between alters, and that this reported estimate will be the actual geodesic distances between the alters in the underlying social networks. There are optimistic studies which suggest that people can perceive groups and can distinguish members according to their position in a social network (Freeman, Freeman, & Michaelson, 1988). There has also been some research and geodesic distances can correlate well

with the perceived distances (Krackhardt & Kilduff, 1999). However, there are also extensive counter arguments in the literature which suggest that informants' reports are prone to error and inaccuracy (Bernard et al., 1979; Butts, 2003; Krackhardt, 1987). To further complicate matters, human perception on dyadic proximity may vary according to the position of the person who is observing the relation (Wilson, O'Leary, Metiu, & Jett, 2008; Hamilton & Sherman, 1996). These concerns imply that the results we present here should be taken as an optimistic evaluation. Further experiments are necessary to evaluate the impact of informant inaccuracy (Krackhardt, 1987; Siciliano, Yenigun, & Ertan, 2012).

In this work we have taken into consideration that respondents can not recognize every person in the social network, by implementing a recognition number parameter which captures the number of individuals that each respondent can recall. This assumption that all respondents are able recognize an equal number of people is also problematic. Some prior work suggests that individuals exhibit considerable variation of the numbers of alters that they can recognize (Bell, Belli-McQueen, & Haider, 2007; Casciaro, 1998; Hill & Dunbar, 2003). Future extensions of our work should consider modeling this variation in recognition number.

In implementing RDS, we assume that the respondents distribute their coupons to individuals with whom they have closest ties. In the present implementation the new RDS recruits are selected from the neighbors of previous respondents who have been given coupons. While there is evidence that recruits are more likely to be chosen from those who have close ties with respondents (Salganik & Heckathorn, 2004; Wejnert, 2010), we note that recruits may distribute their RDS coupons to people who are not within their ego network. In the future this will also be incorporated in our study.

Our results are based on networks which are generated according to the Barabasi- Albert model. We select the Barabasi-Albert model, because it is simple

to implement, and it is a realistic model that has been validated by numerous researchers, and shown to reflect the topology and degree distribution of the actual social networks (Schneeberger et al., 2004; Dombrowski et al., 2013). However, it may be the case that the Barabasi-Albert model is not suitable for certain types of social networks (Bearman, Moody, & Stovel, 2004). In the future studies, we should also run our simulations in networks generated with models other than Barabasi-Albert, as well as using actual social network data previously collected by other researchers.

Our finding are results of simulations, and they demonstrate that it is possible to estimate social network topology with our approach. Although the simulation is a cost-effective method to test an approach before implementing, we confess that our experiments is away from covering problems that can occur during human interaction. Therefore, in future studies, we should also conduct small-scale field studies to evaluate our approach.

Finally, our approach may lead to large numbers of objects to be shown in interviews. As previously mentioned, classical approaches such as use of a roster may be impractical for very large number of objects. In addition, large number of objects may lead to very time-consuming interviews. However, we believe that the social proximity perceptions of the respondents can be efficiently and quickly elicited by using technologies such as interactive software and touchscreens. Future studies should, therefore, address this design and implementation issue.

9 Conclusions

SNA is challenging, and especially is so when the network under consideration is large, when the community is hard-to-reach, or when there is a lack of reliable data.

Hackers, child-porn users, customers and deliverers of online black markets

are hidden networks of cyberspace which exploit the anonymous nature of internet to conduct their criminal acts. Researchers wishing to study such networks need rigorous research tools and methods rather than classical approaches.

In this thesis, we develop a new data collection technique and associated analysis methods to harvest information about the topological structure of social networks where there is no prior information available, and where the members are concerned about their privacy and anonymity.

Through simulation experiments, we test our novel data collection approach. In contrast with prior techniques, we do not ask interviewees to disclose their own ties, opting instead to ask them only to report on perceived dyadic distances among pairs of prior respondents.

Our results indicate that respondents' perceptions of perceived network distance between pairs of alters can be aggregated efficiently to produce estimates of social network distances with high accuracy.

We observe that network size, recognition ratio, and perceivable proximity threshold have significant effects on the performance of our models. We find that larger networks require more interviews to reach the same level of accuracy when compared to the requirements for smaller networks. The recognition ratio, which controls the number of objects that are shown in each interview is observed to have a positive impact on correlation coefficient values. Increasing the recognition ratio enhances the rate at which models converge to their maximal correlation values over the interview process. Similarly, when the perceivable proximity threshold is increased, we observe higher correlation values are achieved earlier on the interview process.

We also find out that sampling methods have an impact on the performance of our new data collection scheme. In particular, Respondent Driven Sampling significantly outperforms random sampling both in terms of node discovery and

correlation values. We find that selecting objects which were closer to the interviewee improves the rate at which a model can attain its high correlation values in the course of interview process.

Our schemes are efficient because they permit respondents to report perceived distances between pairs of alters even when the intermediary nodes on the geodesic connecting these alters are not part of the studied sample.

10 References

Apache comman math library. (2009). Retrieved 2014-05-26, from http://commons.apache.org/proper/commons-math/download_math.cgi

Arsovska, J. (2012, January). Researching difficult populations: Interviewing techniques and methodological issues in face-to-face interviews in the study of organized crime. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 397–415). Springer New York. Retrieved 2014-07-17, from http://link.springer.com/chapter/10.1007/978-1-4614-3876-2_23

Barabasi, A. L., & Albert, R. (1999, October). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. Retrieved 2014-05-19, from http://www.sciencemag.org.ez.lib.jjay.cuny.edu/content/286/5439/509 (PMID: 10521342) doi: 10.1126/science.286.5439.509

Bearman, P. S., Moody, J., & Stovel, K. (2004, July). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, *110*(1), 44–91. Retrieved 2014-07-25, from http://www.jstor.org/stable/10.1086/386272 doi: 10.1086/ajs.2004.110.issue-1

Bell, D. C., Belli-McQueen, B., & Haider, A. (2007, May). Partner naming and forgetting: Recall of network members. *Social Networks*, *29*(2), 279–299. Retrieved 2014-02-16, from http://www.ncbi.nlm.nih.gov.ez.lib.jjay.cuny.edu/pmc/articles/PMC2031835/ doi: 10.1016/j.socnet.2006.12.004

Bernard, H. R., Killworth, P. D., & Sailer, L. (1979). Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, *2*(3), 191–218. Retrieved 2014-06-20, from http://www.sciencedirect.com/science/article/pii/0378873379900145

doi: 10.1016/0378-8733(79)90014-5

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods and Research*, *10*, 141-163.

Butts, C. T. (2003). Network inference, error and informant (in) accuracy: a bayesian approach. *Social Networks*, *25*, 103–140.

Butts, C. T. (2008, March). Social network analysis: A methodological introduction. *Asian Journal of Social Psychology*, *11*(1), 13–41. Retrieved 2014-06-20, from `http://onlinelibrary.wiley.com/doi/10.1111/j.1467-839X.2007.00241.x/abstract` doi: 10.1111/j.1467-839X.2007.00241.x

Casciaro, T. (1998, October). Seeing things clearly: social structure, personality, and accuracy in social network perception. *Social Networks*, *20*(4), 331–351. Retrieved 2014-02-16, from `http://www.sciencedirect.com/science/article/pii/S0378873398000082` doi: 10.1016/S0378-8733(98)00008-2

Coleman, J., Katz, E., & Menzel, H. (1957, December). The diffusion of an innovation among physicians. *Sociometry*, *20*(4), 253–270. Retrieved 2014-06-20, from `http://www.jstor.org/stable/2785979` doi: 10.2307/2785979

Dombrowski, K. (2012). Estimating the size of the methamphetamine-using population in new york city using network sampling techniques. *Advances in Applied Sociology*, *02*(04), 245–252. Retrieved 2014-02-16, from `http://www.scirp.org/journal/PaperDownload.aspx?DOI=10.4236/aasoci.2012.24032` doi: 10.4236/aasoci.2012.24032

Dombrowski, K., Curtis, R., Friedman, S., & Khan, B. (2013, January). Topological and historical considerations for infectious disease transmission among injecting drug users in bushwick, brooklyn (USA). *Sociology Department, Faculty Publications*. Retrieved from

http://digitalcommons.unl.edu/sociologyfacpub/229

Dombrowski, K., Khan, B., Moses, J., Channell, E., & Dombrowski, N. (2012). Network sampling of social divisions in a rural inuit community. *Identities*(0), 1–18. Retrieved 2014-02-16, from http://www.tandfonline.com/doi/abs/10.1080/1070289X.2013.854718 doi: 10.1080/1070289X.2013.854718

Freeman, L. C., Freeman, S. C., & Michaelson, A. G. (1988). On human social intelligence. *Social and Biological Structures*, *11*(4), 415–425.

Gansner, E., Koutsofios, E., & S., N. (2006). Drawing graphs with dot. *Dot File Manual*. Retrieved from http://www.graphviz.org/Documentation/dotguide.pdf

Hamilton, D. L., & Sherman, S. J. (1996, April). Perceiving persons and groups. *Psychological Review*, *103*(2), 336. Retrieved 2014-03-13, from http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=9606162171&site=ehost-live

Heckathorn, D. D. (1997, May). Respondent-driven sampling: A new approach to the study of hidden populations. *Social Problems*, *44*(2), 174–199. Retrieved 2014-02-16, from http://www.jstor.org/stable/3096941 doi: 10.2307/3096941

Heckathorn, D. D. (2002, February). Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. *Social Problems*, *49*(1), 11–34. Retrieved 2014-02-16, from http://www.jstor.org/stable/10.1525/sp.2002.49.1.11 doi: 10.1525/sp.2002.49.1.11

Hendricks, V., & Blanken, P. (1992). Snowball sampling: theoreti- cal and practical considerations. in snowball sampling: A pilot study in cocaine use. *IVO, Rotterdam*, 17–35.

Hill, R. A., & Dunbar, R. I. M. (2003, March). Social network size in humans. *Human Nature*, *14*(1), 53–72. Retrieved 2014-03-11, from http://link.springer.com/article/10.1007/s12110-003-1016-y doi: 10.1007/s12110-003-1016-y

Holt, T. J., Strumsky, D., Smirnova, O., & Kilger, M. (2012). Examining the social networks of malware writers and hackers. *International Journal of Cyber Criminology*, *6*(1), 891–903.

Jones, C., & Volpe, E. H. (2011, April). Organizational identification: Extending our understanding of social identities through social networks. *Journal of Organizational Behavior*, *32*(3), 413–434. Retrieved 2014-06-20, from http://onlinelibrary.wiley.com/doi/10.1002/job.694/abstract doi: 10.1002/job.694

Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, *9*, 109–134.

Krackhardt, D., & Kilduff, M. (1999). Whether close or far: Social distance effects on perceived balance in friendship networks. *Journal of Personality and Social Psychology*, *76*(5), 770–782. doi: 10.1037/0022-3514.76.5.770

Lu, Y., Polgar, M., Luo, X., & Cao, Y. (2010). Social network analysis of a criminal hacker community. *Journal of Computer Information Systems*, *51*(2), 31–41.

Marsden, P. (1990, January). Network data and measurement. *Annual Review of Sociology*, *16*, 435–463. Retrieved 2014-06-20, from http://www.jstor.org/stable/2083277

Marsden, P. (2005). Recent developments in network measurement. In P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis* (pp. 8–30). New York: Cambridge.

McCormick, T. H. (2011). Statistical methods for indirectly observed network data. *Unpublished Ph.D Thesis*. Retrieved 2014-06-26, from http://academiccommons.columbia.edu/catalog/ac:131447

McGloin, J. M., & Kirk, D. S. (2010). An overview of social network analysis. *Journal of Criminal Justice Education*, *21*(2), 169–181. Retrieved 2014-06-26, from http://dx.doi.org/10.1080/10511251003693694 doi: 10.1080/10511251003693694

McNeeley, S. (2012, January). Sensitive issues in surveys: Reducing refusals while increasing reliability and quality of responses to sensitive survey items. In L. Gideon (Ed.), *Handbook of survey methodology for the social sciences* (pp. 377–396). Springer New York. Retrieved 2014-07-17, from http://link.springer.com/chapter/10.1007/978-1-4614-3876-2_22

Meyer, I. H., & Wilson, P. A. (2009). Sampling lesbian, gay, and bisexual populations. *Journal of Counseling Psychology*, *56*(1), 23–31. doi: 10.1037/a0014587

Miller, K. W., Wilder, L. B., Stillman, F. A., & Becker, D. M. (1997, April). The feasibility of a street-intercept survey method in an african-american community. *American Journal of Public Health*, *87*(4), 655–658. Retrieved 2014-02-16, from http://ajph.aphapublications.org.ez.lib.jjay.cuny.edu/doi/abs/10.2105/AJPH.87.4.655 doi: 10.2105/AJPH.87.4.655

Muhib, F. B., Lin, L. S., Stueve, A., Miller, R. L., Ford, W. L., Johnson, W. D., ... Team, C. I. T. f. Y. S. (2001). A venue-based method for sampling hard-to-reach populations. *Public Health Reports*, *116*(Suppl 1), 216. Retrieved 2014-02-16, from /pmc/articles/PMC1913675/?report=abstract (PMID: 11889287)

Oracle. (2014). *Java se development kit 7u60.* Retrieved 2014-06-13, from http://www.oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html

Papachristos, A. (2011). The coming of a networked criminology?, using social network analysis in the study of crime and deviance. In J. MacDonald (Ed.), *Measuring crime and criminality* (Vol. 13, pp. 101–140). New Brunswisck, NJ: Advances in Criminological Theory.

Parker, J. G., & Asher, S. R. (1993). Friendship and friendship quality in middle childhood: Links with peer group acceptance and feelings of loneliness and social dissatisfaction. *Developmental Psychology*, *29*(4), 611–621. doi: 10.1037/0012-1649.29.4.611

Petersen, R. (2005). Using snowball-based methods in hidden populations to generate a randomized community sample of gang-affiliated adolescents. *Youth Violence and Juvenile Justice*, *3*(2), 151–167. Retrieved 2014-02-16, from zotero://attachment/191/151.full_3.html doi: 10.1177/1541204004273316

Ramirez-Valles, J., Heckathorn, D. D., Vazquez, R., Diaz, R. M., & Campbell, R. T. (2005, December). From networks to populations: The development and application of respondent-driven sampling among IDUs and latino gay men. *AIDS and Behavior*, *9*(4), 387–402. Retrieved 2014-02-16, from http://link.springer.com/article/10.1007/s10461-005-9012-3 doi: 10.1007/s10461-005-9012-3

Salganik, M. J., & Heckathorn, D. D. (2004, December). Sampling and estimation in hidden populations using RespondentâĂŘDriven sampling. *Sociological Methodology*, *34*(1), 193–240. Retrieved 2014-02-16, from http://onlinelibrary.wiley.com/doi/10.1111/j.0081-1750.2004.00152.x/abstract doi: 10.1111/j.0081-1750.2004.00152.x

Schneeberger, A., Mercer, C. H., Gregson, S. A., Ferguson, N. M., Nyamukapa, C. A., & Anderson, G. P., R. M.and Garnett. (2004). *Scale-free networks and*

sexually transmitted diseases: A des... : Sexually transmitted diseases. Retrieved 2014-06-28, from http://journals.lww.com.ez.lib.jjay.cuny.edu/stdjournal/Fulltext/2004/06000/Scale_Free_Networks_and_Sexually_Transmitted.12.aspx

Scott, P. (2000). Handling relational data. In *Social network analysis: A handbook 2000* (pp. 39–62). CA: SAGE Publications.

Siciliano, M. D., Yenigun, D., & Ertan, G. (2012, October). Estimating network structure via random sampling: Cognitive social structures and the adaptive threshold method. *Social Networks*, *34*(4), 585–600. Retrieved 2014-02-16, from http://www.sciencedirect.com/science/article/pii/S0378873312000408 doi: 10.1016/j.socnet.2012.06.004

Spreen, M. (1992). Rare populations, hidden populations, and link-tracing designs: What and why? *Bulletin de MÃlthodologie Sociologique*, *36*(1), 34-58. Retrieved 2014-06-19, from http://bms.sagepub.com/content/36/1/34 doi: 10.1177/075910639203600103

Stueve, A., O'Donnell, L. N., Duran, R., Doval, A. S., & Blome, J. (2001, June). Time-space sampling in minority communities: results with young latino men who have sex with men. *American Journal of Public Health*, *91*(6), 922. Retrieved 2014-02-16, from /pmc/articles/PMC1446469/?report=abstract (PMID: 11392935)

Sudman, S. (1976). *Applied sampling.* Academic Press, New York.

Wasserman, S., & Faust, K. (1994a). Social network analysis in social and behavioral sciences. In *Social network data collection and application* (p. 1-66). USA: Cambridge University Press.

Wasserman, S., & Faust, K. (1994b). Social network analysis, method and applications. In *Social network data collection and application* (p. 29-29). USA: Cambridge University Press.

Watters, J. K., & Biernacki, P. (1989). Targeted sampling: Options for the study of hidden populations. *Social Problems*, *36*(4), 416-430. Retrieved 2014-03-13, from http://www.jstor.org/stable/800824 doi: 10.2307/800824

Wejnert, C. (2009, August). An empirical test of respondent-driven sampling: Point estimates, variance, degree measures, and out-of-equilibrium data. *Sociological Methodology*, *39*(1), 73–116. Retrieved 2014-06-21, from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9531.2009.01216.x/abstract doi: 10.1111/j.1467-9531.2009.01216.x

Wejnert, C. (2010, May). Social network analysis with respondent-driven sampling data: A study of racial integration on campus. *Social Networks*, *32*(2), 112–124. Retrieved 2014-02-16, from http://linkinghub.elsevier.com/retrieve/pii/S0378873309000501 doi: 10.1016/j.socnet.2009.09.002

Wilson, J. M., O'Leary, M. B., Metiu, A., & Jett, Q. R. (2008, July). Perceived proximity in virtual work: Explaining the paradox of far but close. *Organization Studies*, *29*(7), 979–1002. Retrieved 2014-02-16, from http://oss.sagepub.com/content/29/7/979 doi: 10.1177/0170840607083105

Appendix

Class Diagrams

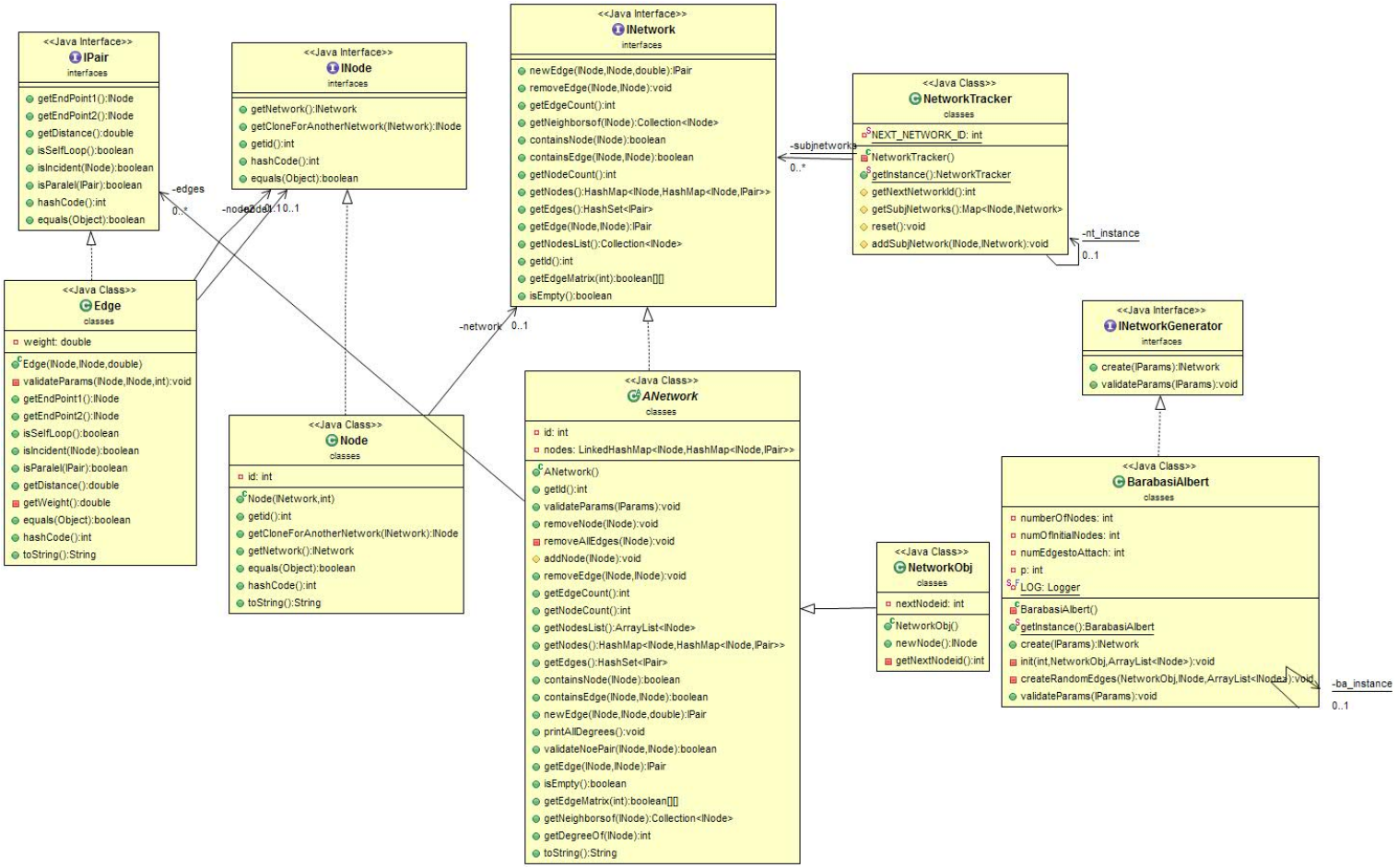


Figure A1. This class diagram illustrates the custom classes used in generating reference graph.

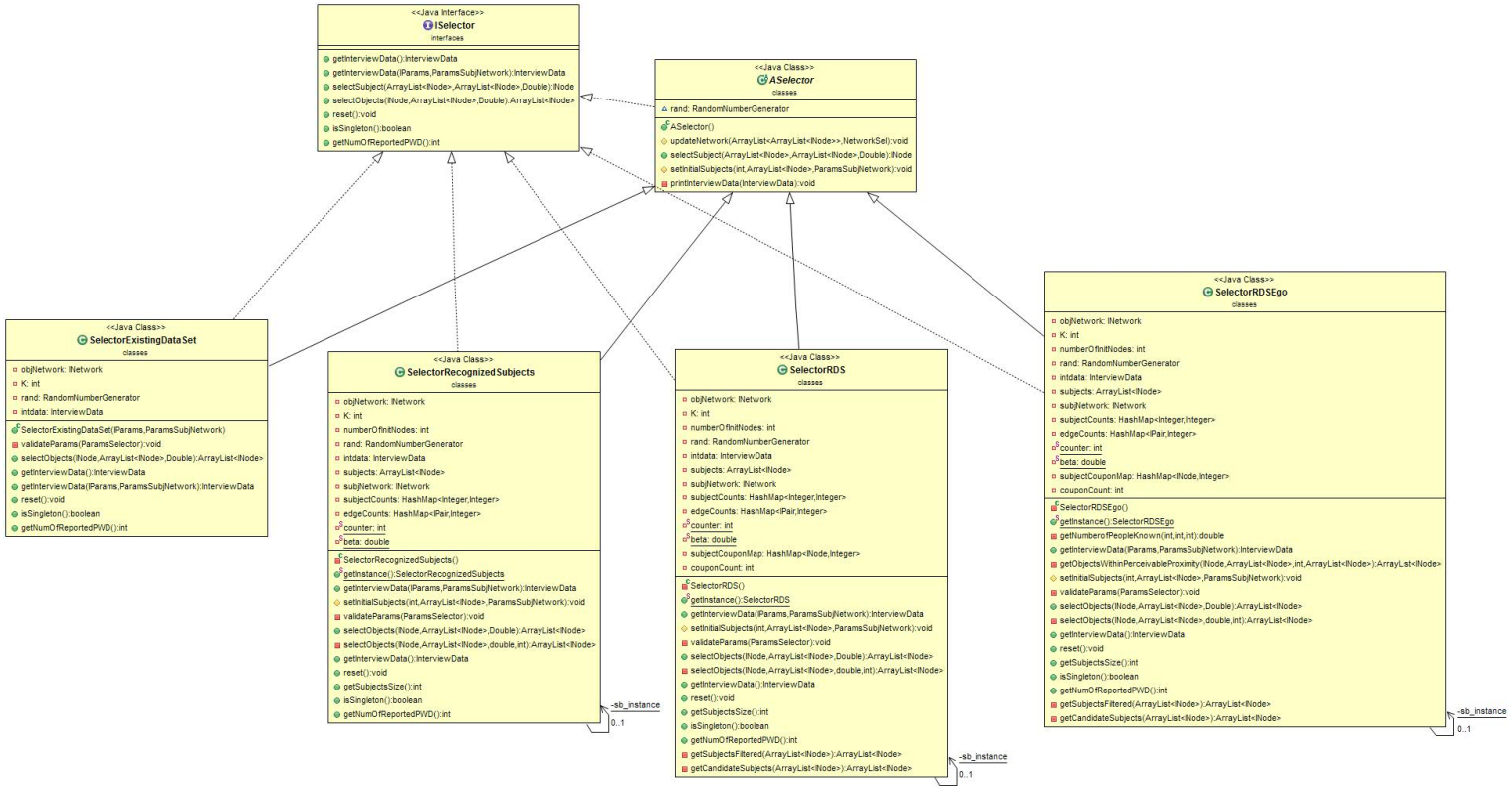


Figure A2. This class diagram illustrates the custom classes used in different selection scheme.

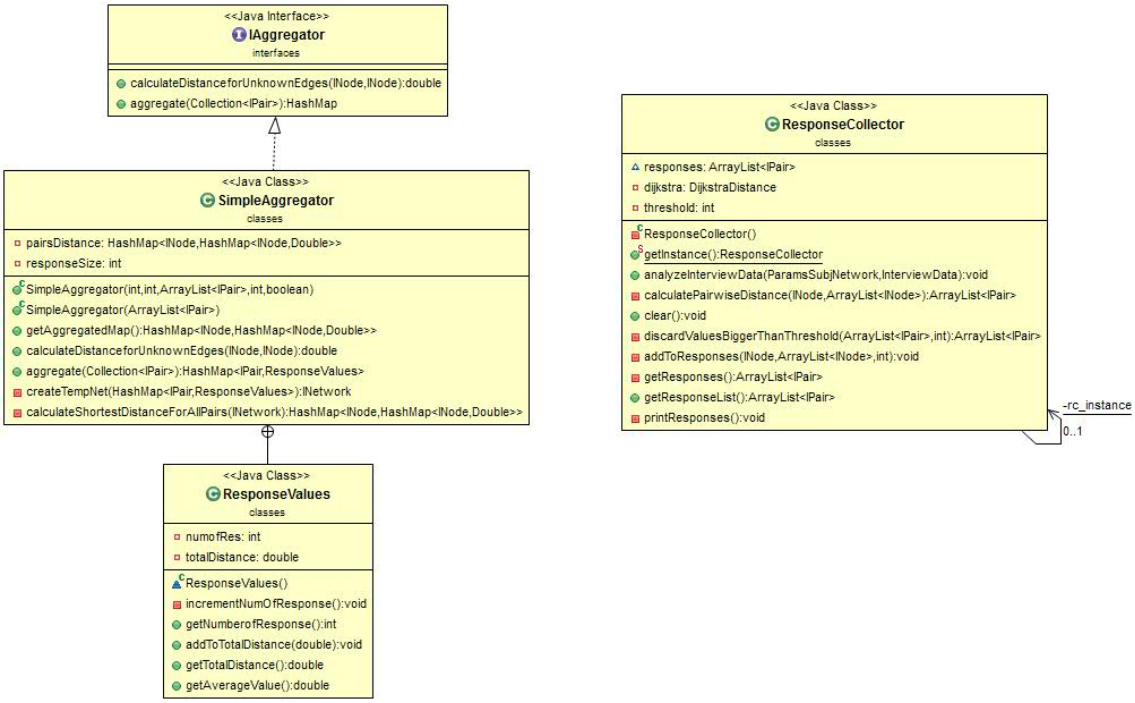


Figure A3. This class diagram illustrates the custom classes used in aggregation

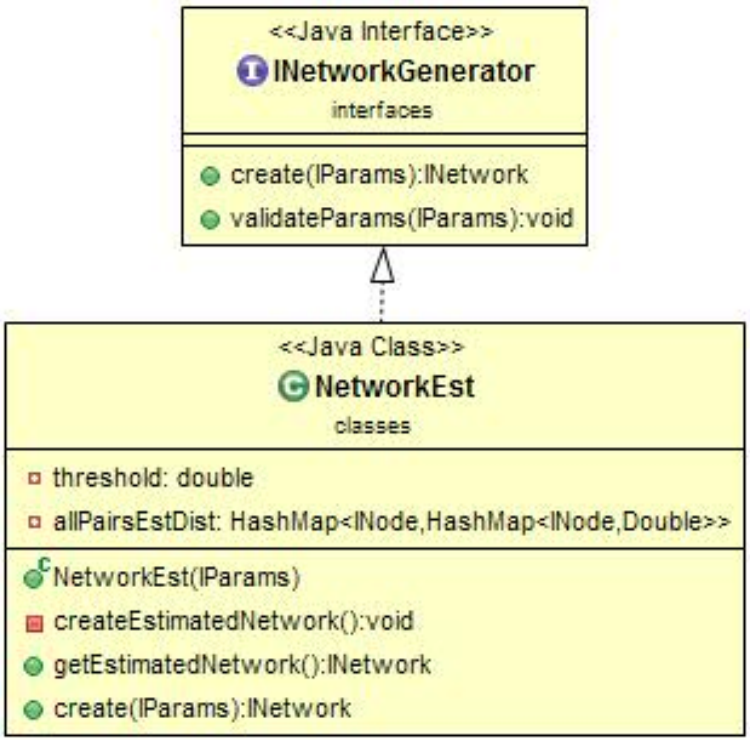


Figure A4. This class diagram illustrates the custom classes used in generating estimated graph.

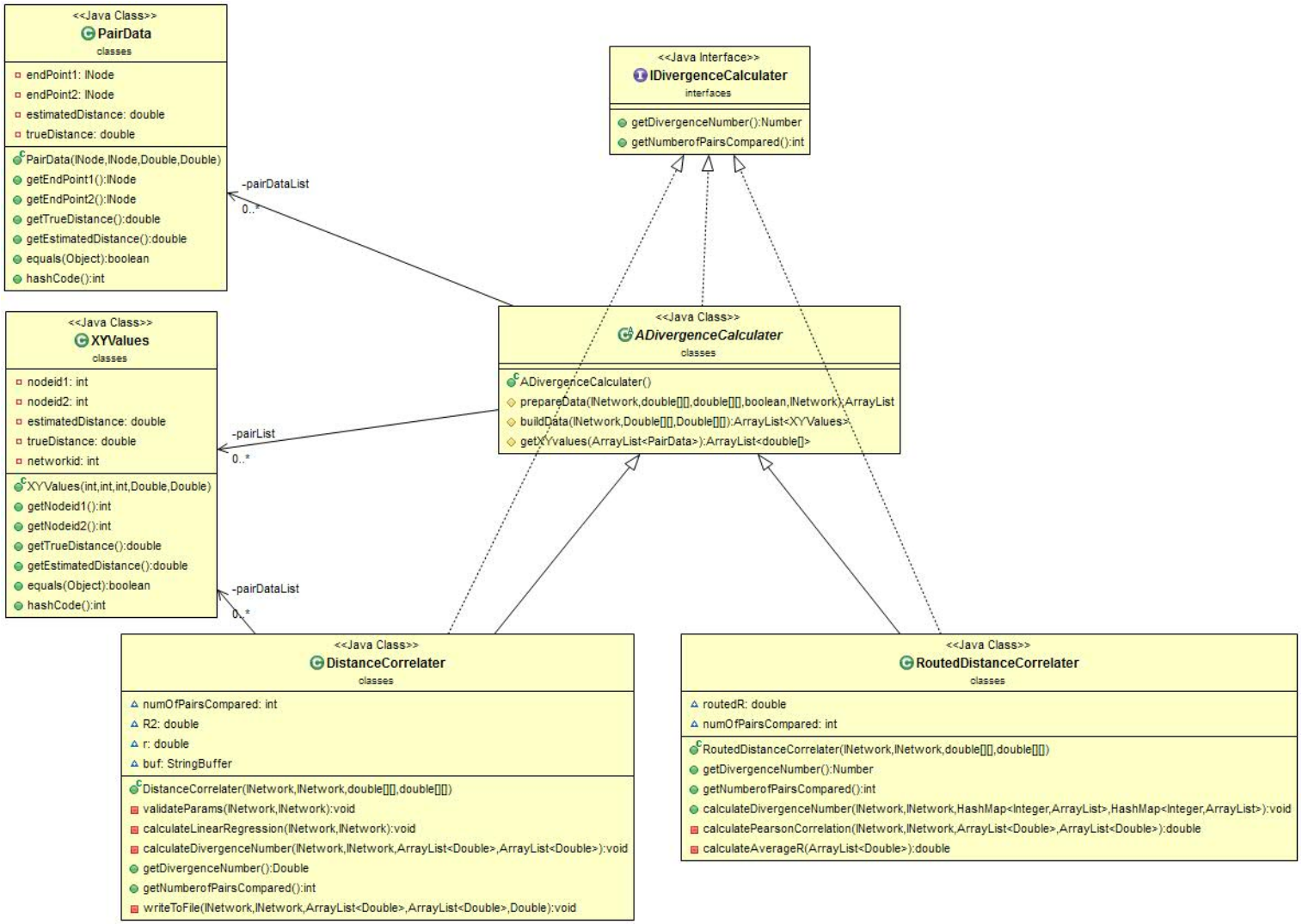


Figure A5. This class diagram illustrates the custom classes used in different evaluation methods.