Perspectives on social network analysis for observational scientific data

Lisa Singh and Elisa Bienenstock and Janet Mann

Abstract This chapter is a conceptual look at data quality issues that arise during scientific observations and their impact on social network analysis. We provide examples of the many types of incompleteness, bias and uncertainty that impact the quality of social network data. Our approach is to leverage the insights and experience of observational behavioral scientists familiar with the challenges of making inference when data are not complete, and suggest avenues for extending these to relational data questions. The focus of our discussion is on network data collection using observational methods because they contain high dimensionality, incomplete data, varying degrees of observational certainty, and potential observer bias. However, the problems and recommendations identified here exist in many other domains, including online social networks, cell phone networks, covert networks, and disease transmission networks.

1 Introduction

Social network analysis (SNA) is an empirical methodology for formally describing the structure of relationships between observed entities. Historically, the use of social network metrics to identify and distinguish socially relevant information about group structure and the relationships among group members was validated by empirical work on small closed systems. The metrics supported the intuition of researchers who had a deep knowledge of the social context of individuals in a particular social setting. When used for purely descriptive purposes on data sets with

Lisa Singh
Georgetown University, Washington, DC 20057, e-mail: singh@cs.georgetown.edu

Elisa Bienenstock
Georgetown University, Washington, DC 20057, e-mail: ejb57@georgetown.edu

Janet Mann
Georgetown University, Washington, DC 20057, e-mail: mannj2@georgetown.edu

explicit boundary conditions [43], the metrics proved to be robust and beneficial. In the last decade, with an increase in popularity and the exponential growth in computational power, the uses of SNA have extended beyond description, and include making inferences about context from network structure. While this is an exciting direction for SNA, a number of questions arise - an important one being: how do we determine when these inferences are reasonable? Even though in some situations reasonable inferences should be possible, many data sets that apply SNA for inferences are incomplete, biased and contain high degrees of uncertainty.

In this chapter, we provide examples of the many types of incompleteness, bias and uncertainty that impact the quality of social network data. Our approach is to leverage the insights and experience of observational behavioral scientists familiar with the challenges of making inference when data are not complete, and suggest avenues for extending these to relational data questions. Rather than discuss data collection challenges generally, this paper is focused on one example, a 25 year long study of wild dolphins, which includes numerous examples of incompleteness, uncertainty and bias, but which also provides ample data to develop and test techniques for compensating for these problems and improving the quality of inferences drawn about network structure. While our longterm goal is to develop computational approaches to compensate for different data quality issues that arise, this chapter is a conceptual look at data quality issues that arise during scientific observations and their impact on social network analysis. The focus of our discussion is on observational scientific networks; however, the problems identified here exist in many other domains including online social networks, cell phone networks, covert networks, and disease transmission networks.

This chapter is organized as follows. In the next section, we introduce definitions and background describing the missing data problems. In sections 3, 4 and 5, we provide examples of each of these problems in the context of a large bottlenose dolphin data set and present recommendations for improving data quality for SNA. Section 6 briefly describes some open areas of research for computer scientists. Conclusions are presented in section 7.

2 Definitions and background

There is a misconception that SNA is a statistical method [25]. Descriptive statistics and SNA are similar as they provide metrics to characterize patterns in observed data. They differ with regard to their ability to generalize and provide inference beyond the sample measured. Inferential statistics allows researchers to generalize beyond the data observed provided the data meet certain criteria. One principle assumption is that observations are independent, an assumption at odds with the purpose of social network analysis, which is to capture the dependencies between observations. As a result, direct application of classical statistics may lead to incorrect inferences. While approaches for building statistical models for interrelated entities do exist [44], a consensus approach that adequately considers the interrelated population or addresses inference has yet to emerge. A second assumption is that the data sampled represent the population. Statistical solutions address random error, not systamatic bias. There is no recognized methodology used to identify or correct for sampling choices that result in samples with bias, which is often the case with social network data. Therefore, it is important to understand how robust SNA metrics are to different types of data incompleteness, uncertainty, and bias for both descriptive and predictive models.

Observational scientists, particularly those that collect field data are familiar with a range of methodological issues concerning imperfect data. While field data are considered to have high ecological or external validity compared to experimental approaches, the degree of external validity can be undermined by sampling or data collection bias. Observational scientists are limited, as observations cannot be universal or persistent. Researchers must make choices about how to sample their subjects efficiently and sufficiently to learn about their behaviors. There are fundamental sources of limitations that must be addressed. While other data quality factors exist, this paper focuses on three: completeness, certainty, and bias.

Our discussion applies to different types of social networks, e.g. one-mode, two-mode, multi-relation, etc. For ease of exposition, we focus on one-mode networks, where the graph $G = (V,E)$. Formally, our network G contains a set of vertices $V = \{v_1, v_2, ...v_n\}$, where n is the number of vertices in the network, and a set of edges between vertices in V: $E = \{e_{ij} = (v_i, v_j) : v_i \textit{ and } v_j \in V\}$. Each vertex may be associated with a set of attributes, while each edge is associated with a set of relationship types and features about those relationships.

Observational scientists collect data and assign observed events to V and E. The challenge for the social network researcher is to get sufficient data to adequately represent the population of individuals (V) and evidence of relationships between these individuals (E) such that the metrics calculated on the available data represents the true population under investigation.

2.1 Completeness

An observational scientist monitors a subject for an interval of time. Example subjects include dolphins, humans, or genomes. Each monitoring period can be viewed as a sampling period consisting of a number of observations. Monitoring periods may include one or more events. Depending on the sampling protocol, events could include the behavior, movements, or interactions of one or more subjects. In an open system, observational scientists can only monitor a subset of the entire population and a subset of the relationships. Having relational information about an entire population is an optimal situation for applying social network analysis.

> The social network population is **known** if V contains every entity of the population within the network. The social network connectivity is **known** if E contains every relationship of the entities in the population within the network. If both V and E are **known**, the social network, G is considered to be *structurally complete*. Otherwise, we say that G is structurally

incomplete. G' is a single sample or subgraph of G. We refer to the set of samples we have of G as $S(G) = \{G'_i | G'_i \subset G\}$.

In other words, if all the individuals in a population are known and all the relationships between the individuals are known, we consider the network derived from the population to be structurally complete. This is sometimes the case in closed systems, but never the case in open systems. Open system challenges differ because of 1) the accessibility of the subjects (monitoring the subject 24 hours a day may not be possible); 2) the flow of individuals in and out of the observation area; and 3) the sampling approach chosen.

2.2 Certainty

In many contexts during an observation, there are distinct types of uncertainties that can arise. We begin by describing uncertainty in the context of social network graph topology. We then consider it in the context of data quality during data collection and redefine some general measures to make them applicable to social networks.

Social network uncertainty: An observer may be uncertain about the number of subjects in a group. For instance, in animal behavioral studies the group members may be moving so rapidly that only a range can be certain, e.g. there were between 15 and 20 dolphins in the group. Or there may be ambiguity about subject identity such that the same individual is counted more than once. This case corresponds to one individual with multiple identities on a social networking site. In relation to SNA, this certainty maps to *node existence certainty*. Related to this is *node identity certainty*, where researchers are uncertain about the identities of the individuals in the group, e.g. there are ten howler monkeys but only six can be clearly identified. A third type of certainty concerns whether or not two animals have an association between them. This corresponds to *edge existence certainty*. Another type of uncertainty is related to attributes of edges or nodes. *Node attribute certainty* focuses on certainty of any node feature, e.g. sex, individual behavior, etc. *Edge attribute certainty* maps to uncertainty related to features about associations, e.g. association behavior (the animals are playing). All of these types of certainty directly influence the overall reliability of the data for both descriptive and predictive tasks. Further, as more observations of the same nodes, edges, and attribute are taken, each individual observation is further validated and the overall certainty of those observed values increases. In this way, as sample size increases, certainty and data reliability also increases.

Data quality and uncertainty: Singh et al [38] present the following quality measures associated with data collection: observation certainty, observation detail consistency, researcher vocabulary confidence and data stability. We redefine the first three in the context of social networks. Observational certainty is defined as the degree of confidence in the measurement itself [38]. In the context of social networks, it is the probability that both of the individuals and a mutual behavior were observed.

For each observation event, OE, the certainty of the event is the probability that the event occurred, $p(OE)$. If a social network G' is built for individuals observed during OE, then each node, v_i, and each edge, e_{ij} has a probability of occurrence that represents the certainty of the individuals and relationships in G' respectively, $p(v_i)$, $p(v_j)$, and $p(e_{ij})$.

Knowing whether or not an event occurred and who participated in the event are necessary for accurate inference. If a researcher is making inferences and is not identifying the uncertain components of the network, the results of SNA may be misleading. If too much uncertainty exists in the observation, the inferences based on SNA are unreliable. Therefore, it is important to identify degrees of certainty of data during data collection. Doing so improves the overall understanding about the quality of the data and the accuracy of the resulting inferences.

When conducting statistical analysis on traditional observation data, the potential for uncertainty and error exist for the individual and the behavior observed. However, we see from the previous definitions that the potential for uncertainty and error is higher in social network data because error can also occur when measuring the relationship data. Sources of observational uncertainty for SNA include intermittent or poor observation conditions, distance between subjects, indistinguishable characteristics of subjects, ambiguous behaviors, and behaviors that are not mutually exclusive.

Differences in depth of data collection across researchers are relevant for both completeness and certainty. We refer to this as observation detail consistency. Here we ask if two different scientists are recording the same level of detail during an observation? For example, if one scientist records a red bird and another records a male, undersized cardinal eating a juniper berry, the observation detail consistency will be low. In the context of SNA, the observational detail corresponds to the attribute values associated with each vertex or edge in the network. The larger the number of known attribute values, the higher the observation detail. More precisely, given a set of researchers, observation detail consistency = 1 when every researcher records the same level of detail during observation event OE. It is 0 when every researcher records a different level of detail during the same observation event.

Another quality concern involves identification of a common frame of reference or a consistent language interpretation across researchers in a group. For example, one researcher may look at a rabbit and suggest that it is large. Another researcher looking at the same rabbit may classify it as medium-sized. Singh et al [38] refer to consistent interpretation among data collectors as researcher vocabulary consistency.

2.3 Bias

Bias is present when the likelihood of observation or collection of behavioral data is correlated in some way with a variable of interest to the researcher. This is especially insidious, as it is impossible to measure correlations when data are missing. This unrepresentative sample can result from sampling bias and observer bias, differences in collection criteria among researchers on a project, or different combination of all three.

A recent study illustrates this point: The classically maligned species, the spotted hyena, until recently, was reputed to be a lowly scavenger and kleptoparasite, stealing the prey from the nobler lion. Recent research [42] included nocturnal observations which revealed that the spotted hyenas were the masterful hunters, and it was lions who stole the majority of kills. Though by morning, researchers (and tourists) found lions feeding on the wildebeest while the rightful owners circled.

Daytime sampling was sufficient for answering research questions related to daytime behavior. However, the data sample was insufficient for making more general inferences. If the researcher makes inferences based on incomplete data, then these inferences are biased toward the collected data. For this analysis, data collected only during the day was biased as it could never reveal the nocturnal behavior if it differed from daytime activities, and until observed, these differences could not be measured or accounted for. We pause to point out that this example also highlights limitations with classical statistics. Even though classical statistics methods control for random errors, observational bias is not addressed.

Observer bias is generally defined as the tendency to bias observational or other data toward unconscious or even conscious expectations. Although inter-observer reliability can ameliorate this tendency, it is not uncommon for multiple observers to have the same expectations or implicit hypotheses and hence, the same biases. Most insidious is when observers have the same expectation and thus agree on the outcome without realizing the source of agreement. Even when experimenters or observers are blind to the hypothesis, they often have an implicit premise. Observer bias can also occur because some subjects or behaviors are more obvious or easier to observe or classify than others. Some subjects are more distinctive in appearance (e.g., bottlenose dolphins with mangled dorsal fins), or in behavior (e.g., active socializing more obvious than resting). It is common for observers to, for example, classify a group as 'socializing' when only a minority are doing so (see also Altmann 1974 [2]; Mann 1999 [29]). Attention is naturally drawn to the socializing individuals and it might appear as if most of the group is involved.

Observational scientists are aware of possible biases inherent in many of their design decisions [14], like observing subjects at specific times or based out of specific locations, and have developed strategies to compensate or control for these. In section 5, we present some of their techniques with those used in other communities.

3 Dolphin Societies

We now describe the wild dolphin case that will be considered in the remainder of the chapter. After describing data set and its relevance to human societies, we consider the different challenges described earlier and provides concrete examples to highlight the pitfalls of different observational approaches and present some potential remedies that are important for SNA.

3.1 Shark Bay Data Collection

Researchers have monitored Indian Ocean bottlenose dolphins in Shark Bay, Australia since 1984. This site offers unparalleled conditions for the study of dolphin behavior: clear, shallow water, a high dolphin density, and relatively low human-related disturbance. All researchers use systematic protocols for monitoring individual dolphins and contribute information on identification, births, deaths, weaning, scars, ranging, behavior, association, and basic ecological data. Among the 1200+ individuals studied since 1984, 95% are recognizable by dorsal fin features. Over 470 calves have been tracked from birth. These data have been collected using standard quantitative sampling techniques including point, scan, and continuous sampling [29], and include structured values, textual descriptions, photos, and geospatial data. The complementary sampling strategies are important for SNA since it provides a way to use triangulation to improve the confidence of the SNA and inferences based on the SNA [2].

The Shark Bay data set includes extensive survey data ($14,000$ records) and more intensive focal follow data ($2,800$ hrs). Brief (5-10 min) surveys include records on location, behavior, associates, habitat, photographic information, and physical data. They present a ”snapshot” of association and behavior. Focal follows (1-9 hrs per follow) provide detailed minute-to-minute behavioral information including group composition, activity, location, and specific social interactions using standard quantitative sampling techniques. These focal data provide a detailed depiction of dynamic individual behavior across four generations.

3.2 Fission Fussion Societies

Most dolphin societies are characterized by extreme fission-fusion dynamics where spatial and temporal stability of groups is low [3]. Bottlenose dolphins, like humans, change their associations over the course of minutes to years, but maintain stable relationships within this fluid structure. In fact, one could easily argue that bottlenose dolphins exhibit greater similarity in social structure to humans than most other primates [11]. Like modern humans, dolphins live in large unbounded complex societies in which individuals maintain short and long-term bonds spanning more than 30 years, and have multi-level alliance structures (alliances of alliances, e.g., [11]). They also exhibit similarly slow life-histories such as prolonged maternal care, late age of maturity, and long-lifespan [30].

From a social network perspective it is important to understand when particular individuals associate with others. For instance, activity budgets have obvious implications for social interactions, and in a fission-fusion species, on the degree of sociability. As an example, foraging is inversely related to sociability [18, 19]. Work on dolphin social structures is in its infancy. However, dolphin social structure seems to be amenable to both ego-network and group-level SNA methods (e.g., see Lusseau et al. [28] and Stanton & Mann [20]).

3.3 Advantages and disadvantages of non-human studies

One of the advantages of studying non-human animals is that they can be observed in their natural habitat for long periods compared to humans. Even so, observers have numerous practical and logistical limitations, and cannot view or monitor their subjects 100% of the time. Some have recommended the use of small cameras or other recording technology, but because of the difficulty of using these devices, particularly with bottlenose dolphins, even data collected in this way would produce incomplete, biased networks.

Given the difficulty of tagging, more traditional, human intensive methods are still used to study dolphins, all of which require the researcher to make decisions during observation. It is incumbent on the researcher to identify the uncertainties and biases introduced and attempt to reduce them.

Finally, when observing animals with low travel costs and long distance communication, it is difficult to determine boundaries of groups. Many different criteria used to define bottlenose dolphin groups ranging from the most conservative 10-m chain rule [39], [17] to all dolphins estimated to be within 100-m radius of the boat/observer [35] to all individuals within sight moving in the same general direction, interacting or engaged in similar activities [13]. The group ranging decision has a large impact on the SNA structure since conservative approaches may miss associations, while less conservative approaches may include erroneous associations. Further, since different researchers use different methods for defining groups, comparisons across sites may not always be possible. For the Shark Bay Dolphin Research project, researchers use the 10-m chain rule.

4 Completeness of Network - Sampling subjects and collecting enough data

One primary challenge of an observational researcher is to observe a large range of behaviors and events. Since researchers of open systems cannot view all the subjects all the time, they must learn as much as possible using only a limited window on their subjects. Many factors have to be considered when collecting data including cost and safety of the researchers.

In this section we begin by looking at different sampling options for social network data. We then compare network measures computed on dolphin data illustrating how different sampling approaches can lead to different outcomes for two social network measures, degree and clustering coefficient. Finally, we discuss sample size and present some recommendations related to completeness of a social network.

4.1 Sampling Options

When considering sampling, there are initial questions that must be answered: 1) what sampling approach should one use when collecting observational scientific data for SNA; 2) how does one know that the social network built from the sample is a good structural representation to the actual social network; and 3) how does one know enough data has been collected for the analysis?

A number of researchers have begun working on methodologies for sampling for social network data. Different types of samples include convenience/opportunistic samples, simple random samples, systematic samples, stratified random samples, and cluster random samples, and snowball sampling designs [22, 36]. It is difficult for researchers who observe animals to obtain a completely random sample. Convenience or stratified random samples (based on location or sex for strata) are more readily available. Snowball sampling, is a network focused approach of building a sample network by selecting seed nodes, collecting data on the social network of the seed nodes and then iteratively adding their associates to the data sample [22]. Snowball sampling can be either breadth first or depth first. In both cases, members are added to the sample until the network sample is determined to be sufficient, either because it achieves a pre-specified size or because no new network members are identified. In observational research, because of the high cost associated with keeping track of all the subjects during a snowball sample, quasi-snowball sampling can be a more realistic approach. It limits the number of individuals followed at each stage.

As was mentioned in the last section, the Shark Bay project uses multiple approaches for sampling subjects, focal sampling (a quasi-snowball sampling approach), general survey sampling (a convenience random sample), and transect sampling (a systematic random sample). Surveys (or sightings) are assessments of a group's or individual's status, typically including demographic, ecological and behavioral data. These can by opportunistic (based on group or individual encounters) or systematic (e.g., transect surveys, searching along a particular pre-defined route). Focal sampling or follows are systematic sampling periods on an individual or in some cases a pair (see Mann 1999 [29]; Altmann 1974 [2]). These have less locational bias but might reduce the variance in the data as there is a greater cost to focusing so intently on a pair of individual. For focal follows, dolphin pairs are selected using a randomized list order. As the season progresses, dolphins that have been followed the least, i.e. the fewest hours, are selected to maintain similar observation durations for the different dolphin pairs.

4.2 Sampling Methods Comparison

Each sampling approach has its strengths and weaknesses and is associated with particular biases. In the context of SNA, previous work demonstrated on synthetic data that social network metrics were robust to missing or erroneous data [9, 15] .

In other words, even when some data are missing, some important features of the network can still be ascertained, e.g. actors with high centrality or high betweenness. These approaches were focused on data missing due to error, not bias. Stanton et al [40] conducted an analysis that compared different social network measures when different sampling strategies were used for sampling the same dolphins in Shark Bay. Table 1 shows some of the results in [40] along with some other statistics about the social network for quasi-snowball sampled data and random point sampled data. Notice the large different in number of nodes and degree. We see that while the mean clustering coefficient is similar, using both complementary approaches give the researcher an opportunity to better approximate the ground truth.

Table 1 Comparing quasi-snowball and random point sampling

Sampling Method	Nbr of Nodes	Mean Normalized Degree	Mean Clustering Coefficient
random point	1000+	33	0.654
quasi-snowball	200+	22	0.653

4.3 Amount of data per subject necessary

Although scientists are often concerned with power or effect size in order to conduct their analysis, they rarely consider how much data per subject is needed. The former typically concerns itself with the number of subjects to include in an analysis or when planning an experimental study, not the amount of data per subject.

There are a number of ways to approach this problem. One is to subsample the entire dataset and analyze it according to these amounts (hours per subject or observations per subject) and see when the outcome variable no longer changes or is no longer correlated with the amount of data. For example, we examined how much focal data are needed per subject before the number of associates is no longer correlated with hours of observation. For the Shark Bay dolphins, the answer was 10 hours. That is, we needed to observe dolphin calves for 10 hours or more before we could be confident that we captured most of their network. However, the same was not true for their mothers. Only when we sampled 16 hours or more on mothers was the relationship not significant.

Most dolphin and whale studies use far less data than this. Table 2 shows the variation in network size, degree and clustering coefficient for the focal follow dolphins [40]. It highlights the fact that 10 minutes to 50 minutes is not sufficient to accurately determine the network of a dolphin or to eliminate the correlation between amount of data and dolphin degree.

Few have examined how much observation time is needed (samples per subject) to adequately capture the size (degree), density or other characteristics of weighted or unweighted networks. This obviously depends on within-individual variation. In

Table 2 Comparing quasi-snowball amount of data

Number of Minutes Observed	Network Size	Mean Normalized Degree	Mean Clustering Coefficient
$\geq$ 10 minutes	184	33	0.674
$\geq$ 30 minutes	157	20	0.665
$\geq$ 50 minutes	144	22	0.653

societies characterized by fission-fusion, such as bottlenose dolphins and humans, such variation across different temporal and spatial scales can be substantial.

4.4 Recommendations

Make sure you have enough data
Having limited amounts of data can severely impact the quality of the social network. One way to verify whether the amount of data is ’enough’ is to use a smaller sample with lots of data (as a kind of ground-truth) and compare that sample to a larger sample with less data. If the resulting properties are similar, the sample size is reasonable. The assumption here is that having more data about a small set of subjects is a more reliable, higher quality set of data than the overall data set. If that is not the case, this approach will not lead to a valid analysis. Another approach is to weight data by the amount so that subjects with more data do not count more in the analysis. This is a variation of stratified sampling. Finally, one can also statistically control for the amount of data so that it is a factor in the analysis. Traditional ways of estimating necessary sample size are power analysis, information model comparisons, effect size statistics and Bayesian statistics. It is important to mention that while having more data increases the completeness and certainty of the data, it does not necessarily reduce the amount of bias in the sample. Other strategies presented in the next section need to be considered for that.

5 Identifying uncertainties and biases

As previously discussed, there are multiple layers of certainty and different types of bias. Understanding how confident an observer is about the individual or behavior seen is important for understanding the level of reliability and amount of bias associated with an analysis. Using the Shark Bay data, this section presents a few examples of uncertainty and bias that occur during field research. We focus on problems that have direct implications for accurate SNA.

5.1 Uncertain Subjects and Behaviors

Figure 1 shows a group of dolphins seen in a survey observation. The blue nodes represent dolphins the researchers are very familiar with an so they have high node identity certainty. The green represents dolphins with moderate node identity certainty. In other words, the researchers code some level of uncertainty in the identity of these nodes. Finally, though the researchers are certain that three more dolphins are present, these can not be identified. The node identity certainty of those three nodes is zero.

There are a few observations this small example highlights. First, maintaining only the certain information would reduce the group size by more than 30%. Second, assuming all the uncertain identities are certain, may lead to conclusions that may or may not be accurate. Finally, keeping track of multiple layers of uncertainty provides more information than maintaining only one level. In other words, it is still informative to know that 19 dolphins are in this group even if the identities of three of them are unknown.

Capturing this level of certainty is also important for attributes and interactions. In this dolphin data set, every interaction and behavior is marked with a certainty level - high, moderate and low. This information can not only be used to understand scientific findings, it can also be used to see how well new researchers are being trained. Using the data, one can determine if there are certain subject behaviors that are difficult for new researchers to capture with high certainty and develop new training approaches or new methodologies for identifying these behaviors.

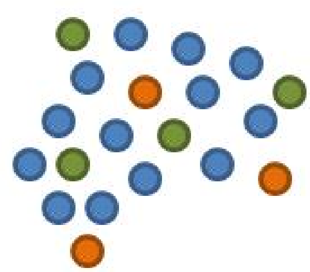

Fig. 1 Dolphins present in a single observation - blue node=high node identity certainty, green nodes=moderate node identity certainty, orange nodes= no node identity certainty.

5.2 Observers reliability and consistency

A number of observer factors can lead to uncertainty and bias in samples. These include observer experience, fatigue, interest or focus, and how behavior or groups are defined. Between observer reliability can sometimes be assessed by using both live coding or videotaped data to compare consistency. This is complicated by a number of factors, including different vantage points for live observations (e.g., on a boat), and that observers may be more vigilant and thus concurrent in their coding when they themselves are being monitored. In addition, observers vary in their experience and it would be difficult for a new field observer to measure up to one with a decade or more of experience.

There are ways around these particular problems. Instead of a single observer, multiple observers can be present to help determine behavior categories. Photographing and videotaping can help verify individuals and behaviors. One can also statistically compare across observers to see what biases they may have introduced into sampling. For example, in the course of analyzing adult female sighting records, it became obvious that one observer at our field site never sighted a dolphin commonly seen by others. The dolphin's dorsal fin was not terribly distinct, so the researcher's team failed to identify her correctly. Researchers were able to go back through the unidentified dolphins in his film records and fill in the missing records. In other cases, the sampling error cannot be eliminated, but can be accounted for statistically or analyzed at another level. For example, some observers might be adequate for distinguishing between foraging and socializing, but cannot correctly identify behavior subtypes (type of foraging or type of socializing). One can either collapse those data into larger categories, or create confidence intervals around the subtypes.

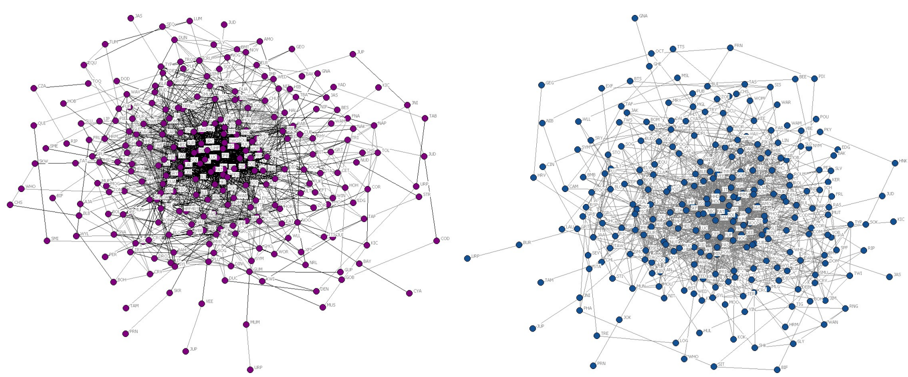

Fig. 2 Different views of the sample network by different researchers. The left network is one researcher's view. The right network is another researcher's view

Figure 2 shows a dolphin social network generated using over 10 years of survey data from two different, experienced researchers on the Shark Bay Research Project. These networks contains only relationships that were viewed at least five times by the researchers. On the surface, the structure of the two networks seems very similar. However, the actual dolphins in the network vary considerably. The second observer did not capture 72% of the relationships captured by the first observer. The first observer did not capture 59% of the relationships captured by the second. More specifically, of the 1145 relationships in the left network, 819 of those relationships were not identified by the second researcher (right network). Similarly, of the 789 relationships observed by the second researcher (right network), 463 were not observed by the first researcher (left network).

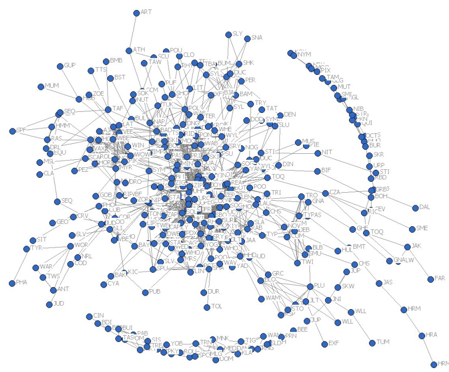

Fig. 3 A network representing the common view of the two researchers, i.e. the associations both researchers observed

There were only 326 associations that were observed by both researchers. Figure 3 illustrates the common network found by both researchers. Notice the large difference in density between the network in Figure 3 and the other two networks in Figure 2. So why is there such a discrepancy? Ultimately, each researcher has certain biases or questions of interest that drive their observations. In this case, one researcher favors larger groups and tended to ignore smaller groups. Also, one of these researchers was very interested in mothers and calves, while the other was interested in male alliances. Their different observation preferences and research objectives may have contributed to the biases in the data collection.

5.3 Time and behavioral sampling

One fundamental difference between humans and dolphins is that humans are mostly diurnal while dolphins and whales are cathemeral, active for periods of the day and night. At present, most data on dolphins was collected during the day, so researchers can only characterize dolphin diurnal behavior. One reason for a preference for daytime observations is that it is easier for human researchers to see during the day. However, being limited by visual access to our subjects can, as discussed earlier, lead to mischaracterizations of species (e.g., spotted hyenas) or of specific individuals.

As an example, one very sociable female named Square (for the notch in her fin) spent only 3% of her day hunting, while 30% is the average for all adult females. Her foraging budget was tiny but stable over a 12 year period. One evening we followed her from 4pm to 10pm, and, sure enough, as soon as the sun set, she hunted for several hours on her own, completely uncharacteristic of her sociable and restful daytime behavior. Here is a case where her day behavior differs from her night time behavior. Since foraging and socialization have an inverse relationship, her day time social network is likely to be different from her night time network.

5.4 Depth and association sampling

Fortunately, dolphins do not change their associates at depth. That is, one can account for associates because their diving bouts are brief, allowing observers to monitor group members spatially. This is not true for sperm whales, who stay submerged for an hour or more and can travel many kilometers during one dive. Although sperm whales live in stable matrilineal units and tend to disperse temporarily during foraging bouts, it would be difficult to determine associations during foraging. In sum, association data collected for some species are more likely to be biased by activity than others. It depends on the temporal scale of sampling. In studies of killer whales, the definitions of group or pod differs depending on the researcher, making comparisons difficult even within the same population. For example, some studies used the visual range of observers and an assessment of acting coordinated as a definition for ”group” [4, 5], while others define associates as those in the same photographic frame in at least 50% of the photos over several years. Those associations are used to define pod membership [7]. It would be difficult to determine relative bond strength for killer whales with such disparate methods for defining associations.

One area where depth does have an impact when studying marine mammals is hunting. Marine mammals typically hunt at depth, out of the observers’ view. It is easier for humans to observe behaviors that occur near the ocean surface, although clearly much of what dolphins do occurs below the surface. Recording only what happens at the surface is a recognized source of bias. Some researchers try to mitigate this by inferring behavior: although the animals are out of view during a dive, if the observer is confident that the behavior is foraging, then it can be recorded as such. Because the behavior is not seen, to reduce the possible bias the information can be annotated with additional information about the degree of confidence or certainty of each behavior and the quality of the observation (at/near surface, subsurface). Direction changes, the type of dive and observations of prey catches can also be used to help the observer confirm the behavior that occurred mostly out of view.

5.5 Hidden behaviors or social encounters

Some behaviors or social encounters are harder to detect because they are only performed when the animal is ”cryptic.” Dolphins and whales do not ”hide” from research observers while foraging or performing other behaviors. However, observers must concern themselves that some behaviors may occur out of view because the animals are hiding from conspecifics, predators or prey. In many primate species, subadult or subordinate males only attempt to mate with females out of view from adults or more dominant males. The consequences of mating in the open can be severe for both sexes. Focal sampling is one of the best ways to address this problem because the observer stays with the individual focal animal regardless of where the animal goes. Even though this method usually works and provides information on a broad range of behaviors, the potential for observational bias persists as observers are more likely to lose their focal when the focal is trying to avoid conspecifics. In Shark Bay, females often avoid adult males and try to do so quietly. Researchers will sometimes follow a focal female and the entire group will dive and seemingly disappear when a group of males is nearby.

5.6 Observers can affect the behaviors they monitor

Most field researchers take years to habituate their population to human presence. This is particularly difficult if the population has been subject to hunting or poaching. In addition, their presence could affect predation rates. Cheetah are known to inhibit hunting in the presence of humans. Vehicles and boat engines make noise that could interfere with communication and can attract or repel predators or prey. Many contraptions that we consider essential for research, such as cameras, make sounds or reflect light that can be disturbing to others.

In Shark Bay, the dolphins became habituated initially through benign interaction with fishers and tourists, and subsequently, by researchers themselves. We use small boats with quiet 4-stroke engines. We check the noise levels underwater (snapping shrimp are much louder than our boats) and we never enter the water with the dolphins. How do we know they are habituated? They show very high tolerance of our presence. They engage in intimate behaviors very close to the boat (nursing, petting, social play), and we often hang back for periods to be certain their movements are not dictated by our presence.

Sometimes a new behavior occurs because of the presence of humans. At many field sites, one behavior, bow-riding, occurs because we are there. Although we discourage it during any sampling periods, there are times we encourage this behavior because dolphins often turn belly-up during bow-rides, which allows us to determine their sex and even age by degree of speckling [39]. Of concern is that some individuals might avoid an approaching boat or vehicle long before sampling begins. In fission-fusion societies this is more likely than in stable groups. Our main strategy for reducing this type of bias is to try and assess group size and behavior from hundreds of meters away. We can usually determine if animals have left or if they change their general behavior before we begin sampling. If the departed animals continue to be evasive, we note this and try and determine identity by photos taken at a distance. In our experience, even these dolphins become habituated to our boats quickly by associating with other habituated animals.

5.7 Recommendations

Improve observation certainty and consistency through standardization
The more detailed the recording of the subject and the event, the closer the observation is to reality. Accurately measuring the discrepancy between the observation and reality is a difficult problem. One way to decrease the discrepancy is to develop surveys that specify the minimum amount of information needed from the observation. Because the survey is created by research participants, it is considered a reasonable approximation of reality and represents a meaningful set of data as perceived by a group of scientists. While it cannot be considered a complete reflection of reality, it does identify important features and helps standardize the level of detail across researchers.

If researchers have developed a survey, one approach to measuring the amount of consistency is to have each researcher use the survey to capture measurements about the same event. Then these surveys can be compared to improve observation certainty for newer researchers and observation detail consistency for all observers.

Mark your biases
If researchers are aware of biases, they can devise sampling and observational design strategies to mitigate these. One important initial step is to identify all the sources of uncertainty and bias that are apparent. For each measurement that has uncertainty, an attribute can be added to the data set identifying the level of confidence in the measurement. Subjective measurements can also be separated from more standard measurements. For example, observation location should be a separate attribute, not placed in a large notes field. A similar strategy can be employed to minimize biases. By marking uncertainty and bias, a researcher can better control for these issues during analysis.

Minimize observer bias
There are several ways to minimize observer bias. First, good quantitative sampling techniques are fundamental. In sampling group behavior, explicit scan sampling, where the behavior of each individual is accounted for (or a randomized subset), is essential. Second, in addition to a clear systematic sampling regimen, a detailed, explicit ethogram is also needed. The ethogram defines each behavior in as much detail as possible while explaining what behaviors cannot co-occur. Third, if behaviors or individuals are ambiguous, record the data and also record a certainty level of the data. It is important to allow for all possible permutations of uncertainty in the data coding. By allowing observers to code for uncertainty, more accurate coding is likely to occur. It will discourage observers from being either overconfident in their

coding, or neglecting to report what they do know. Finally, there is no substitute for making implicit hypotheses as explicit as possible. After all, the job of the scientist is to undermine the hypothesis, rather than prove it.

Compensate for bias
As with any comparative network analysis, having a reliable ground truth is important for the analysis. Using multiple sampling techniques and finding common parts of the network across samples can help identify a more accurate ground truth. More importantly, triangulation is an important strategy for identifying and adjusting for discrepancies and biases in one or more of the samples.

6 Computational Approaches to Improve Data Quality for Social Network Analysis

This chapter has identified different issues that can arise during 1) data collection and 2) application of social network analysis on incomplete, uncertain or biased data. Now we briefly describe a number of contributions that computer scientists can make to help improve the quality of the data and the resulting analysis conducted by social scientists and other researchers.

Develop interactive approaches for data exploration of social networks
Because there are so many factors associated with data exploration, new ways of interacting with data need to be devised. With decades of effort, improved technologies for remote tracking, and heightened concern for disappearing species, the next generation of biologists is measuring more features on larger populations. Novel approaches for data exploration are necessary to better understand different population dynamics. Interactive visualizations can be used to find outliers, highlight uncertainties, and detect possible biases.

Visual analysis of social networks is an integral component of the social network field. Tools fall into two categories, those that have sophisticated statistical analysis using matrix operations and those that focus on interactive visualization of uni-mode networks and multi-mode, heterogeneous networks. Various toolkits have been developed to help programmers create interactive visualizations themselves. The most robust include JUNG [34], Prefuse [24], and GUESS [1]. While tools for heterogeneous networks are emerging [37, 26], a need still exists to develop more tools that incorporate context specific graph visualizations, switch between different granularities of data, handle large data sets, and incorporate sophisticated longitudinal visual analytics.

Develop algorithms that consider uncertain data
Many traditional data mining algorithms ignore uncertainty in attribute and relational data. Developing approaches for clustering, community detection, and anomaly detection that consider the certainty of the attribute and relationships during analysis will result in more accurate approaches for attaining inductive knowledge about social structures. One approach is to represent the uncertainty as probabilities and build probabilistic models for each of the mentioned tasks for networked data.

Use reliable attribute and relational data to determine missing data Different imputation strategies exist for traditional sampled attribute data. These need to be extended to handle missing data in relational data. One approach for doing this would consider multiple subsamples of the original sample and create a ground truth distribution of the entire sample based on the distribution of these subsamples. This ground truth could then be used to help with imputation of missing values.

Predict structural properties networks Newman [32] surveyed approaches for analyzing structural properties of networks. Much of the work has been descriptive in nature, but recently there has been more work which uses structural properties for prediction. Within this category, a number of papers focus on the spread of influence through the network [8, 12, 27]. These papers attempt to identify the most influential nodes in the network. While these approaches are an important start, much work remains to accurately infer structural properties, particularly in the presence of bias.

Identify changing dynamics of social networks While static network analysis is important for understanding network ambiguity, information transmission, network pruning and the network as a whole, it is also important to consider the evolution of network relations over time. To date methods for this typically involve comparing the structures of static social networks at multiple time points (e.g., GEE regression models).

In hidden community identification, researchers attempt to identify subsets of actors that are densely connected to each other, but less densely connected to others. These densely connected regions are called communities. Communities are found using different measures for cohesion and modularity [21, 23, 33, 41]. Other previous research has focused on community detection [33], extraction of unknown community structures [10], simplifying network topology through K-cores [31], and block modeling [6].

The majority of community detection work focused on static networks and constrained the problem to allow an actor to belong to only a single community. Computer scientists have recently begun analyzing the dynamics of social networks, communities and groups [41, 16]. Currently, many of the models make assumptions that limit overlapping group membership and make strict assumptions about the changing dynamics of the participants in the network. The models also ignore sample size. This is a concern since in dynamic analysis, samples at each time point must be large enough for analysis. In a static analysis, there may be 100 observations of an animal. However, in a dynamic one, there may be time periods with only one or two observations. The variation of sample size at different times makes the network less stable. Techniques to compensate for this need to be proposed. We also need to develop strategies to handle the volume of data when multiple relations are combined, community overlap exists and the dynamics of these groups and communities are changing.

Of course many of the mentioned computational challenges are further challenged by the volume of network data that exists. Scalable algorithms and approximation heuristics will need to be considered as data sizes grow.

7 Final Thoughts

In this paper, we bring together some techniques and lessons learned from different empirical traditions, social science, animal behavior, and computer science. Our goal was to present issues related to uncertainty, sampling, data quality and SNA that exist across disciplines. Ultimately, limitations in our ability to observe and collect data on social interactions can have a significant impact on our understanding of social structures. By taking the time to understand what is missing, we have a more clear view of what is known.

References

1. E. Adar. Guess: a language and interface for graph exploration. In *SIGCHI conference on Human Factors in computing systems*, 2006.
2. J. Altmann. Observational study of behaviour: Sampling methods. *Behaviour*, 49:227–267, 1974.
3. F. Aureli, C. Schaffner, C. Boesch, and et al. Fission-fusion dynamics: new research frameworks. *Current Anthropology*, 48:627–654, 2008.
4. R. Baird and L. Dill. Ecological and social determinants of group size in transient killer whales. *Behavioural Ecology*, 7:408–416, 1996.
5. R. Baird and H. Whitehead. Social organization of mammal-eating killer whales: Group stability and dispersal patterns. *Canadian Journal of Zoology*, 78:2096–2015, 2000.
6. V. Batageli. Notes on blockmodeling. *Social Networks*, 19:143–155, 1997.
7. M. Bigg, P. Olesiuk, G. Ellis, and J. Ford. Social organization and genealogy of killer whales (orcinus orca) in the coastal waters of british columbia and washington state. *Reports to the International Whaling Commission, Special Issue*, 12:383–405, 1990.
8. M. Boguna and R. Pastor-Satorras. Epidemic spreading in correlated complex networks. *Physical Review*, E 66 4, 2002.
9. Borgatti, Carley, and Krackhardt. Robustness of centrality measures under conditions of imperfect data. *Social Networks*, 28:124–136, 2006.
10. D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Community mining from multi-relational networks. *In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2005.
11. R. Connor, R. Wells, J. Mann, and A. Read. The bottlenose dolphin, /tursiops/ sp.: Social relationships in a fission-fusion society. *In J. Mann, R. Connor, P. Tyack, and H. Whitehead (Eds.), Cetacean Societies: Field studies of dolphins and whales. Chicago: The University of Chicago Press*, pages 91–126, 2000.
12. P. Domingos and M. Richardson. Mining the network value of customers. *In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.
13. R. W. E.C.G. Owen and S. Hofmann. Ranging and association patterns of paired and unpaired adult male atlantic bottlenose dolphins. *Canadian Journal of Zoology*, 80:2072–2089, 2002.
14. D. Franks, R. James, J. Noble, and G. Ruxton. A foundation for developing methodology for social network sampling. *Behavioral Ecology and Sociobiology*, 63:1079–1088, 2009.
15. Frantz and Carley. Relating network topology to the robustness of centrality measures. Technical Report CASOS Technical Report CMU-ISRI-05-117, Carnegie Mellon University, 2005.
16. L. Friedland and D. Jensen. Finding tribes: identifying close-knit individuals from employment patterns. In *KDD*, 2007.

17. S. Gero, L. Bejder, H. Whitehead, J. Mann, and R. Connor. Behaviorally specific preferred associations in bottlenose dolphins, tursiops spp. *Canadian Journal of Zoology*, 83:1566–1573, 2005.
18. Q. Gibson and J. Mann. Early social development in wild bottlenose dolphins: Sex differences, individual variation, and maternal influence. *Animal Behaviour*, 76:375–387, 2008.
19. Q. Gibson and J. Mann. The size, composition, and function of wild bottlenose dolphin (tursiops sp.) mother-calf groups in shark bay, australia. *Animal Behaviour*, 76:389–405, 2008.
20. Q. Gibson and J. Mann. Individual variation in the sociality of wild bottlenose dolphin mothers and calves in shark bay, australia: Sex differences, early development and maternal influence. *Animal Behaviour*, In Press.
21. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.
22. L. Goodman. Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170, 1961.
23. Habiba, T. Y. Berger-Wolf, Y. Yu, and J. Saia. Finding spread blockers in dynamic networks. In *SNA-KDD*, 2008.
24. J. Heer, S. K. Card, and J. A. Landay. prefuse: a toolkit for interactive information visualization. In *SIGCHI conference on Human factors in computing systems*, pages 421–430, New York, NY, USA, 2005. ACM Press.
25. R. James, D. Croft, and J. Krause. Potential banana skins in animal social network analysis. *Behavioral Ecology and Sociobiology*, 63:989–997, 2009.
26. H. Kang, L. Getoor, and L. Singh. Visual analysis of dynamic group membership in temporal social networks. *SIGKDD Explor. Newsl.*, 9(2):13–21, 2007.
27. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.
28. D. Lusseau, B. Wilson, P. Hammond, K. Grellier, J. Durban, K. Parsons, and et al. Quantifying the influence of sociality on population structure in bottlenose dolphins. *Journal of Animal Ecology*, 75:14–24, 2006.
29. J. Mann. Behavioral sampling methods for cetaceans: A review and critique. *Marine Mammal Science*, 15:102–122, 1999.
30. J. Mann, R. C. Connor, L. M. Barre, and M. R. Heithaus. Female reproductive success in bottlenose dolphins (/tursiops/ sp.): Life history, habitat, provisioning, and group size effects. *Behavioral Ecology*, 11:210–219, 2000.
31. J. Moody and D. R. White. Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):103–127, 2003.
32. M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
33. M. Newman. Detecting community structure in networks. *Eur. Phys. J. B*, 38:321–330, 2004.
34. J. O'Madadhain, D. Fisher, P. Smyth, and et al. Analysis and visualization of network data using jung. *Journal of Statistical Software*, 2005.
35. E. Quintana-Rizzo and R. Wells. Resighting and association patterns of bottlenose dolphins (tursiops truncatus) in the cedar keys, florida: Insights into social organization. *Canadian Journal of Zoology*, 79:447–456, 2001.
36. M. Salganik and D. Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological Methodology*, 34:193–239, 2004.
37. L. Singh, M. Beard, L. Getoor, and M. B. Blake. Visual mining of multi-modal social networks at different abstraction levels. In *Proceedings of the 11th International Conference Information Visualization*, pages 672–679, 2007.
38. L. Singh, G. Nelson, J. Mann, A. K. Coakes, E. Krzyszczyk, and E. Herman. Data cleansing and transformation of observational scientific data. In *In the ACM SIGMOD Workshop on Information Quality in Information Systems(IQIS)*, 2006.
39. R. Smolker, A. Richards, R. Connors, and J. Pepper. Sex differences in patterns of association among indian ocean bottlenose dolphins. *Behaviour*, 123:38–69, 1992.

40. M. Stanton, J. Mann, Q. Gibson, B. Sargeant, L. Bejder, and L. Singh. How much does method matter? a comparison of social networks of bottlenose dolphins (tursiops sp.) in shark bay, australia. In *18th Biennial Conference on the Biology of Marine Mammals*, 2009.
41. C. Tantipathananandh, T. Berger-Wolf, and D. Kempe. A framework for community identification in dynamic social networks. In *KDD*, 2007.
42. M. Trinkel and G. Kastberger. Competitive interactions between spotted hyenas and lions in the etosha national. *African Journal of Ecology*, 43:220–224, 2005.
43. A. Wasserman and K. Faust. *Social network analysis: methods and applications*. Cambridge University Press, Cambridge, 1994.
44. S. Wassermann and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61:401–425, 1996.