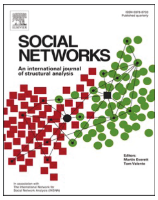
Social Networks 66 (2021) 1–9

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet



Revisiting the accuracy problem in network analysis using a unique dataset

Steven R. Corman [a,*], Elena Steiner [a], Jeffrey D. Proulx [b], Arindam Dutta [a], Alex Yahja [b], M. Scott Poole [b], Visar Berisha [a], Daniel W. Bliss Bliss [a]

[a] Arizona State University, United States
[b] University of Illinois at Urbana-Champaign, United States

ARTICLE INFO

Keywords:
Networks
Accuracy
Perceived Communication
Observable Communication
Situational
Bias

ABSTRACT

A series of papers published by Bernard and colleagues in the late 1970s and early 1980s, dubbed the “accuracy studies,” called into the question the validity of self-reported perceived communication in the study of networks, showing that such reports explain only about 20 % of the variance in directly observed communication. Questions remain about how well the kinds of organizations studied reflect typical formal organizations, the studies’ short observation periods, and manual observation methods. This study revisits the accuracy studies using a unique dataset comprising 144 weeks of network surveys and machine classification of 7,000 h of audio recordings to measure observable communication in a software engineering unit employing 54 people. Results show that correlations between perceived and observed communication over the weeks studied have a lower average than that reported in the accuracy studies but vary considerably from week to week. It also replicates results of earlier research showing that participants tend to overreport communication when they perceive a strong structural relationship to the alters they are rating. This study solidifies our knowledge about network self-reports using a stronger data foundation than prior research employed. Its results, along with the previous research, suggest that perceived communication is not so much a flawed measure of observable communication as it is a related, yet distinct phenomenon. This highlights the need for developments in theory and modeling that articulate the relationship between perceived and observed communication, and Network Reticulation Theory is suggested as a viable approach.

Interest in networks has exploded in recent yearslargely due to the rise of social media and the availability of graph data associated with it. However; the study of networks has a long history; dating back to at least the middle of the 20thcentury; in the social sciences in general and in organizational studies in particular. Studies in this genre describe the structure of communication relationships in social and organizational contexts as a way of discovering how information flows; how one’s position in the network affects resources available; how organizational processes work; and how they can be improved

Communication networks in social/organizational contexts are typically measured with sociometric surveys. In such surveys, participants estimate their communication with others using counts, Likert-type scales or ranks. Researchers initially assumed that these self-reports were an accurate estimate of communication that occurred. Then beginning in the late 1970s, Bernard and colleagues published a series of studies (Bernard et al., 1984, 1982, 1979; Bernard and Killworth, 1977; Killworth and Bernard, 1976, 1979; Killworth et al., 2006) calling this assumption into question.

The “accuracy studies,” as they have come to be known, administered sociometric surveys to participants to measure their perceived communication with others (using either ranks or scales). They then employed three primary schemes for objectively observing communication in organizational contexts. In the first, they used unusual situations in which participants’ communication was logged, including hearing-impaired TTY users, HAM radio operators, and users of an electronic collaboration system (EIES). All three datasets observe mediated rather than face-to-face communication. The first two organizations are informal and leaderless. As described in the study, membership in EIES groups is self-initiated and leadership is informal as well. None of these organizations can be unambiguously mapped to Mintzberg’s (1989) formal organization types, and the researchers admitted that they might be considered “exotic” (Bernard et al., 1982, p. 35). In the second, they used non-participant observation in naturalistic organizational settings (a fraternity, an office, and a tech company) in which

[*] Corresponding author.
E-mail address: steve.corman@asu.edu (S.R. Corman).

https://doi.org/10.1016/j.socnet.2020.12.010

Available online 21 January 2021

Table 1
Summary of correlation/accuracy results from the accuracy studies.

Author(s)	Accuracy or perceived/ observed correlations	Method of measuring observed communication
Killworth and Bernard (1976)	42 % of participants could rank their first communicant first, second, third or fourth. Average correlation (per Bernard and Killworth, 1977) was $r = 0.523$	Logs of communication between hearing impaired TTY users
Bernard and Killworth (1977)	Average correlation across four contexts was $r = 0.382$	Logs of hearing-impaired TTY users and HAM radio operators; walk-through observation of an office and tech firm.
Bernard et al. (1979)	At the clique level, cognitive data differs 160 % from the behavioral clique structure it was intended to represent.	Walk-through observations of a fraternity
Bernard et al. (1982)	Percent accuracy was 36 %–64 %	Logs of the Electronic Information Exchange System (EIES) at the New Jersey Institute of Technology
Killworth et al. (2006)	Correlation between actual and conceptual paths was r = 0.50	Actual shortest paths in the network

an observer walked through the setting every 15 min for a period of days and recorded cases where people were observed interacting. In the third, Killworth et al. (2006) measured a network with traditional sociometric surveys and then asked participants to estimate the first step in a small-world chain, given a particular network member as a target.

They compared the perceived and observed measures, or in some cases structures derived from dyadic linkages (like cliques or shortest paths). They reported either the percentage of cases that agreed or correlations between the two measures. Their measures and results are summarized in Table 1. The correlations between perceived and observed communication ranged from $0.14 \leq r \leq 0.58$ with a mean of $r = 0.45$ across studies. In studies where percent agreement was measured, that value ranged from 36 % to 64 %. If perceived and observed communication behaved like different measures of the same underlying construct, we would expect correlations greater than 0.70 (similar to the correlations between items loading on the same factor in scaling studies), and higher levels of percentage agreement than the 50 % observed in previous studies. Clearly, based on these results, it is not reasonable to assume that perceived communication is an accurate estimator of communication that took place. Bernard et al. (1984) put it more strongly, claiming that "what people say about their communications bears no useful resemblance to their behavior" (p. 499).

There is evidence that factors other than actual memories of interaction influence the perception of network links. Using a laboratory-simulated organization, Corman and Bradford, 1993 tested contextual factors that could contribute to inaccuracy. They found that a perceived social relationship with the group (measured as the number of others with whom a participant reported a relationship) was correlated with commission errors (overreporting communication with others), $r = -0.26$, $p < 0.05$, and communication load (number of speaking turns observed for a participant in one session) correlated with omission errors (under-reporting communication with others), $r = 0.79$, $p < 0.05$. Johnson and Miller (1986) found that participants' perceptions of network connections had a moderate relationship to objective measures of networks that indicated coresidence and exchange relationships. Interestingly coresidence and exchange networks were not highly related to each other.

Research outside the context of organizational networks also concludes that people are not good at recalling their behaviors. Boase and Ling (2013) reported significant correlations in the range $0.23 \leq r \leq 0.74$ between self-reported and logged telephone and SMS use. Singh and Jain (2017), studying similar call data, reported correlations in the range $0.07 \leq r \leq 0.69$. Menon (1993) reported correlations between actual and reported frequency of several behaviors in the range $0.13 \leq r \leq 0.93$. Kobayashi and Boase (2012), using an app to log voice, SMS, and Gmail activity of Android phone users, found correlations in the range $0.04 \leq r \leq 0.48$ between logged and self-reported behavior. They also found that participants overreported communication in general. Brewer (2000) concluded that "across a variety of relations, people forget a substantial proportion of their social contacts when asked to recall them. Even studies with relatively weak test–retest designs show noteworthy levels of forgetting" (p. 40). In a review of literature, Schwartz (1990) agreed, concluding that "respondents will usually base their answers on some fragmented recall from which they attempt to infer a plausible estimate using various inference strategies" (p. 116). Thus, studies in other behavioral contexts indicate that self-reported behavior consistently explains a small amount of variance in observed behavior, like that reported in the accuracy studies.

Critics of the accuracy studies (Burt and Bittner, 1981; Freeman and Romney, 1987; Freeman et al., 1987; Kashy and Kenny, 1990; Kimball Romney and Weller, 1984; Romney and Faust, 1982; Webster, 1992) analyzed and compared structures derived from perceived and observed communication data using techniques like analysis of structural equivalence and nonmetric multidimensional scaling. They found that structures derived from perceived measures exhibit stronger correlations with similar structures derived from the observed measures, explaining around 50 % of the variance. Based on this finding they concluded that self-reports are an acceptable form of data when the objective is to study structural characteristics of networks, even if there are errors in the individual reports. Corman and Bradford, 1993 pointed out that this is a dispute between *methodological individualism*, which favors explanation of social phenomena via characteristics and behaviors of people making up social groups, and *methodological holism*, which studies social structure via emergent properties of collections of individuals.

We argue that the critics' response, while helpful in showing how useful information can be recovered from perceptual data, does not resolve the accuracy issue. First, much organizational network research takes the individualism approach and is used to explain how messages flow between specific dyads. Second, individuals make decisions about communicating with others based on their perceptions of relationships with those others (Corman, 1990; Corman and Scott, 1994a; Singh and Jain, 2017), so understanding how their perceptions differ from objective observations and what factors influence these differences—i.e. the *methodological situationalism* approach described by Corman and Bradford, 1993—is important for theorizing how networks grow and change. Third, even using the holism approach, observed communication still explains at best only half the variance in perceived communication. Thus, the accuracy problem is still relevant, notwithstanding the higher association between more abstract structures derived from perceived and observed communication data.

There are methodological criticisms of the accuracy studies as well. The observed communication data from these studies rely either on situations in which participants maintain logs or on manual observation and coding of interaction. The former is an unusual situation and most communication in organizations is not logged. The latter is limited in terms of getting access to perform observations, the observability of interaction when such opportunities are found, and the resources available for doing the observation and coding. In addition, we do not know how valid the observational schemes used in these studies are. Finally, the observations in the more formal organizations were limited in time. The "office" dataset was collected over four days, and the "tech" was collected over one week. The longest observation period was for the "EIES" dataset (a period of 4–5 months) but as already noted that study observed mediated communication in a subnetwork of a larger organization.

To our knowledge, the accuracy problem has never been studied in a typical organizational context over an extended period, using ubiquitous observation. It may well be that there is variation in accuracy even when using the same organization, participants, and observation methods. We also do not know how adequate observational schemes like those used in the accuracy studies are, compared to more ubiquitous observation. Studies like the one reported here are increasingly called for owing to the replication crisis in social science (Shrout and Rodgers, 2018). But there is a theoretical issue here as well: If research shows that perceived networks are not straightforward indicators of observable behavior, then there will be a need to develop and test theory designed to explain how they are different and how they are related.

This study is designed to address shortcomings of the original accuracy studies using rigorous, and more detailed observations than were available when the original accuracy studies were done. It uses a unique dataset described below to test the hypothesis:

H1. Communication as reported by participants is an accurate predictor of observable communication between those participants.

To reflect existing approaches to this hypothesis, we test H1 at both the dyadic and structural levels.

The data we have available also afford the opportunity to evaluate and replicate the findings by Corman and Bradford, 1993 that participants with higher communication load tend to under-report their communication with others, and that participants tend to overreport communication when they perceive a strong structural relationship to the others they are rating. Using the week as the unit for cataloging load, we hypothesize that:

H2. Weeks with higher average load among participants will exhibit lower correlations between perceived and observed communication than weeks with lower average communication load among participants.

The rationale for hypothesis 2 is rooted in the fact that participants' perceptions of the network are shaped by interacting with others and by observing others interact. When communication load is relatively low, participants have more cognitive resources available to keep track of interactions. But as load increases, these cognitive resources are increasingly taxed. Higher communication load can therefore obscure participants' ability to track their own interactions as well as interactions between others in the network.

H2 deals with errors of omission, but there is also the possibility of errors of commission. Based on perceived/observed discrepancies in the overall sample, we predict:

H3. Participants who have stronger structural relationships with others will over report communication with those others, relative to what is observable.

When participants are formally related to one another, as when they are assigned to the same work unit or are in superior-subordinate relationships, they are likely to form expectations that they will communicate. As a result, when asked to recall communication relationships, they are likely to overestimate their communication with other participants who are formally related to them.

Methods

Setting

The setting for this research is the Software Factory (SF), a service unit at a large southwestern university providing software engineering services for funded research projects and university technology spinouts. SF had directors and work was led by a professional software engineer who managed student programmers using industry-standard engineering processes and were organized in forma, project-based teams. These characteristics put it squarely in the category of a professional organization (Mintzberg, 1989). It operated for 144 weeks from late 2002 to early 2005, and had 79 participants, including the manager, employees, clients and researchers. Over this time, SF worked on 31 separate projects, developing applications for the social sciences, natural sciences, and education, and for internal use (such as an activity reporting system). This study used only records from the 54 SF employees, because only employees made entries in a code repository and activity reporting system, data we used to test H2 and H3.

Data collection

In addition to developing applications for external clients, SF was established with another purpose, to support social science research on networks. Employees consented to participate and contributed to regular and ongoing data collection. Whenever in the facility, participants wore portable digital audio recorders fitted with lapel microphones. When the participant logged in, a system turned on their recorder and wrote a time stamp. When participants left the facility, they would connect their recorder to the logging system, which would download and store their recording. Over the study period we collected about 7,000 h of these recordings.

Other data were periodically collected. Participants completed weekly sociometric surveys. They were presented with a list of other participants, and for each would report their frequency of communication with that person over the previous week, using a seven-point Likert scale anchored with "(almost) never" and "(almost) constantly." Other data includes recordings of group meetings, regular interviews with participants, notes from periodic non-participant observation, records from a code repository indicating lines inserted/deleted/changed by specific participants, and records from an activity reporting system that tracked hours spent by employees on various tasks and projects.

Because this study collected sociometric data including voice recordings and other personal data, we designed protocols to ensure informed consent and privacy of participants. These included verbal and written explanations of the study and data to be collected and protections for participants, which were acknowledged in written consent agreements from participants. We assured them we would not share data with any other researchers for a minimum of five years after completion of the study, and after that period we would only release data that could be anonymized. They were also allowed to request destruction of data collected about them within the past month (though no employee ever made such a request). These protocols were approved in a full-board IRB review.

Audio recording analysis

We used a simple speech activity detector combined with inter-recording correlations to build a classifier to detect interaction between participants. The idea behind this method is that if two people are interacting at a normal conversational distance, their voices will appear on both recorders, generating a high correlation between the two audio signals (Corman and Scott, 1994b) when properly time-aligned. We determined a minimum number of audio segments required to establish the validity of the classifier (compared to human raters) by using the confidence interval equation (Neyman, 1937), with an error margin (variance per sample) of 5% and an estimated population proportion of 0.8. Calculations showed that the minimum number of samples required was 64 ten-minute segments.

In our analysis, a total of 75 ten-minute (or 3000 15-second) audio segments from random working days and between random dyads were coded by human raters to develop a "gold-standard" for validation. To establish reliability, two trained coders coded a subset of nine randomly selected 10-minute audio segments. Coder training consisted of a review of the tasks, purpose, and audio detection tool functions. The two coders discussed, refined, and applied coding rules for identifying what sounds did and did not constitute conversation. Krippendorff's alpha for categorical data (Hayes and Krippendorff, 2007) was used to establish reliability. For the full set of audio segments $\alpha = 0.93$, suggesting a high level of inter-rater agreement.

Table 2
Results of the simulation to validate our imputation method.

Trial	Correlation	SD Observed	SD Imputed	SD Difference
1	0.88	1.69	1.57	0.12
2	0.89	1.71	1.63	0.07
3	0.87	1.76	1.61	0.14
4	0.85	1.58	1.47	0.11
5	0.91	1.85	1.75	0.10
6	0.90	1.83	1.73	0.10
7	0.91	1.81	1.67	0.13
8	0.92	1.86	1.81	0.06
9	0.90	1.67	1.58	0.09
10	0.90	1.82	1.73	0.09
Average	0.89	1.76	1.66	0.10

To add rigor to this assessment, we conducted a follow-up analysis to test whether inter-rater agreement was inflated by silence in audio segments. We removed all 15-second segments identified as silence before re-calculating reliability. For this test, $\alpha = 0.85$ suggesting we maintained high inter-rater agreement. After establishing reliability, 66 additional 10-minute segments were divided equally and independently coded by each coder. In total, 75 10-minute audio segments were coded to be used to validate the detection system.

Using this data, we developed a machine classifier using simple speech-feature-based threshold and cross-correlation technique to detect communication, trained on these coded segments. The receiver operating characteristic curve (ROC) was constructed to measure the performance of the classifier, which plots the true positive against the false positive rate (Powers, 2011). The ROC is a probability curve and the area under the curve (AUC) represents the degree of separability. The ROC-AUC of the system was evaluated to be 0.88, which means that the system was able to reproduce the coded data with a probability of 0.88. More technical information about this method is available on request from the authors; see the Appendix for further details on development and testing of the classifier.

After establishing the validity of the classifier, we applied it to our entire recording dataset. We computed for each week, for each pair of participants, the number of minutes they were observed interacting to produce edges of a valued, directed network. The resulting dataset contains 6330 edges, with mean edge weight of 110.01 min (s.d. 124.74).

Survey imputation

Most employees did not have a perfect record of completing the sociometric surveys, leading to gaps in the data. We chose to impute missing values in some of these cases. Due to a lack of payroll records, we do not know if the gaps were due to a participant simply skipping the survey or being away for a time. We assumed that any gaps of more than four weeks were not due to just skipping the survey, and we did not do imputation in these cases. We also did not do imputation of more than one missing survey around the time of Christmas break, since most of the employees were off at this time. For the remaining cases we imputed missing values as the average of the values for the preceding and following surveys for a given participant. The survey data contains 24,862 valued, directed edges, 7934 (31.9 %) of which were imputed. Of the imputations we performed, 63.5 % of the values were in gaps of one week, 24.1 % were in gaps of two weeks, 9.5 % were in gaps of three weeks, and 2.9 % were in gaps of four weeks.

Imputation based on averages such as we have done can reduce variance in the variable of interest. Multiple imputation techniques are typically used to guard against this; however, we know of no existing procedures for multiple imputation of valued networks over time (see Huisman and Steglich, 2008, for a method for binary networks). To determine whether our imputed values are likely to differ significantly from those that would have been observed, we conducted a simulation to compare actual observations with imputed values, computed as if the actual observations were missing. To reflect the distribution of gaps noted above, we randomly selected 320 samples of gap size one, 125 of gap size two, 50 of gap size three, and 14 of gap size four, for a total of n = 720 cases. For each case we randomly selected an ego and alter. Then from all weeks where ego rated alter, we randomly selected a set of consecutive weeks to include a "before" observation, one or more gap observation(s), and an "after" observation. For each case, we computed the average of the before and after values and recorded this value paired with the actual gap observations, then computed the correlation between the actual observations and their imputed values.

Results of ten runs of this simulation are shown in Table 2. The average correlation between the actual observations and their matching imputed values over ten trials is $r = 0.89$ ($r^2 = 0.79$). The average standard deviation of the imputed values is about 5% lower, but given that only 31.9 % of the values were imputed in the actual dataset, we believe this should have a negligible effect on correlations between perceived and observed communication in our main analysis.

Analysis

To analyze the correspondence between the perceived and observed data, we created adjacency matrices for each week for both data types by dividing the edge data for the perceived and observed networks into weekly segments and aggregating edges for the week. To account for possible differences in the way participants used the perceived communication scales, we transformed their ratings into z-scores per participant (i.e., we standardized each participant's ratings based on the mean and standard deviation of their ratings across all alters and administrations). We used these values in a correlation analysis for each week. We included only participant pairs where both survey responses and recording observations were available for the week in question and where there were at least three nodes in the resulting network. This resulted in 120 pairs of weekly networks, with a minimum of three nodes, a maximum of 12, and an average node count of 7.43 (s.d. 2.22).

To test dyad-level accuracy, we analyzed each pair of adjacency matrices with the Quadratic Assignment Procedure (QAP; Krackhardt, 1988) as implemented in UCINET. QAP computes a standard observed correlation between the elements of the two matrices. It then conducts a simulation, randomly permuting the two matrices 1000 times, each time re-computing the correlation to yield a distribution. QAP does not produce a standard parametric significance test; indeed, it cannot because the observations are not independent. Instead it tests the chances of obtaining a larger correlation than the one observed, given multiple random re-orderings of the two adjacency matrices. We considered a QAP correlation to be significant if the probability of obtaining a simulated correlation greater than or equal to the observed correlation was $p \leq 0.05$.

Regarding structural similarity, early accuracy studies (Killworth and Bernard, 1976, 1979) used triad census (Holland and Leinhardt, 1975) to compare patterns of coordination and cohesion between recalled networks and observed networks. Both studies found more transitive triads in perceived networks than in observed networks, and Kilworth and Bernard (1979) report a $r = 0.46$ correlation between perceived and observed triad counts, suggesting some degree of structural similarity. In our study, a triad census would not be a representative measure of structural equivalence. Neither the perceived nor the observed network contain isolates, and both networks have a relatively high graph density, leading to little variance in transitivity. Therefore, a structural measure that incorporates additional sources of variance such as edge weights is better suited to an assessment of structural similarity in our case. In this study, we measure network cohesion using correlation transitivity (Dekker et al., 2017) to compare perceived and observed networks for 121 weeks of the observation period. Correlation transitivity measures the correlation between edge weights (amount of communication) and the proportion of transitive ties in the network

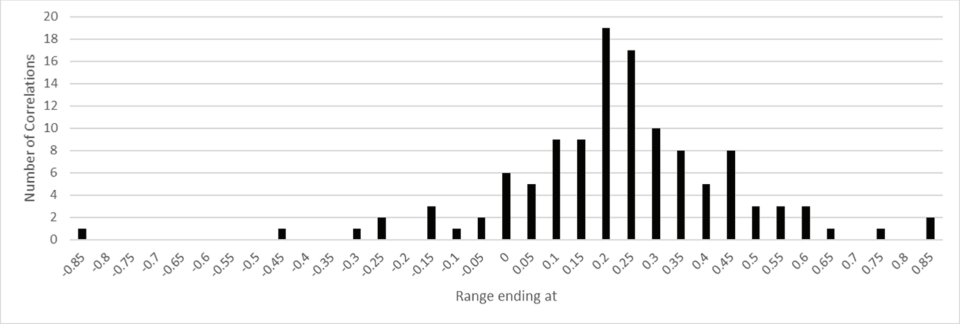


Fig. 1. Distribution of correlations between perceived and observed measurements over all weeks, in r = 0.05 ranges.

(transitive network structure). Like other calculations of transitivity, positive values are interpreted as the network having a greater prevalence of collective coordination, and negative values are interpreted as a tendency towards brokering or gatekeeping.

Critics of the accuracy studies have also assessed the influence of communication volume on structural similarity. For example, Romney and Faust (1982) revisit Bernard et al. (1982) with a follow-up analysis using multidimensional unfolding to assess structural similarity between rank-ordered perceived communication and observed communication. They use adjacency matrix marginals that represent total interaction volume for each network actor to compare perceived and observed networks using Coombs's multidimensional unfolding (Coombs, 1958) to normalize their data and allow for direct comparison. They report a correlation of $r = 0.74$ for their projected data, suggesting that the assessment of structural similarity may add an additional degree of context to the interpretation of accuracy studies.

Romney and Faust had all participants rank their degree of communication with all other participants. In our study we do not have a census of rank data for each participant for each week, making an analysis identical to that of Romney & Faust impossible. However, assessing the degree of similarity between standardized adjacency matrix marginals for communication volume for each week offers a comparable method to assess structural similarity. For 121 weeks of data, we used correlation analysis to assess the degree of similarity between adjacency matrix marginals. Edge weight marginals were standardized to allow for direct comparison across perceived and observed networks.

To test H2 we computed correlations between the QAP correlations and the average communication load (i.e. among participants included in the networks for that week) for that week. We operationalized load two different ways: average hours recorded in the activity reporting system, and average insertions/deletions/changes that the participant made in the code repository (another measure of how much work they did and therefore how much they are likely to communicate). Neither system was in place until the 24th week of operation, so these tests are based on data from 97 weeks of observation.

To test H3, we computed standardized perceived and observed measures for each participant pair and summed the difference between these across the study period. We then computed two different measures of structural relationships for these same pairs. One was the sum over the project of the number of times one participant reported being paired with another for a programming task in the activity reporting system. The second was the one-mode projection (for participants) of the two-mode network linking participants to projects in the activity reporting system. The link weights of this network represent the number of times a pair of participants reported working on the same project.

Results

In the test of H1 at the dyadic level, QAP correlations were significant for 51 (58.1 %) of the 121 weeks included in the analysis. The range of observed correlations was $-0.86 \leq r \leq 0.93$. The average correlation over all the weeks was $r = 0.25$ (SD = 0.27), and for the significant weeks it was $r = 0.32$ (s.d. 0.14). The average variance explained is 6.25 %, and 15.36 %, respectively. The distribution of the values of the correlations within 0.05 ranges is shown in Fig. 1. The bulk are within the range $0.00 \leq r \leq 0.45$.

To test H1 at the structural level, we calculated correlation transitivity for 121 weeks of perceived and observed networks. The relationship between the perceived and the observed networks based on correlation transitivity is $r = 0.29$ ($p < 0.05$) which echoes the findings of previous studies and suggests some degree of structural similarity. We also tested whether correlation transitivity varies to a significant extent between recalled networks and observed networks. The *t*-test shows a significant mean difference between the series (per $M = 0.56$, obs $M = 0.69$, $p < 0.05$) suggesting that the degree of transitive structure is higher in the observed network on average. To check whether measures of correlation transitivity are stationary throughout the observation period, we tested for significant correlations with time. Only the perceived network correlates to the point of significance ($r = .36$, $p < .05$). The partial correlation between perceived and observed networks that controls for time is $r = 0.34$ ($p < 0.05$) suggesting that the relative agreement between the two series is not excessively influenced by time.

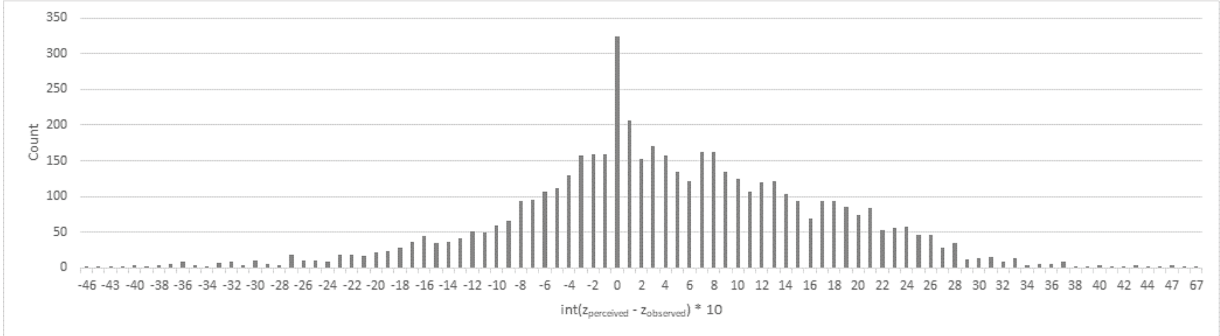


Fig. 2. Distribution of standardized perceived-observed differences over all edges.

Edge weight marginals for communication volume were calculated for 121 weeks of perceived and observed networks. For out-bound ties, the correlation between the perceived and the observed networks is $r = 0.26$, $p < .05$, and for inbound ties is $r = 0.41$, $p < .05$. In follow up, we used partial correlations controlling for time to check for stationarity throughout the observation period. For both perceived and observed networks, inbound and outbound communication volume negatively and significantly correlated with time in the range $-0.26 \leq r \leq -0.16$. The partial correlation between perceived and observed networks ($0.27 \leq r \leq 0.38$) suggests that the relative agreement between these series is not excessively influenced by time. These results, at both the dyadic and structural levels, lead us to reject H1.

To test H2, we computed correlations between two average load measures and the perceived/observed correlations from the QAP tests for each week. Correlation with activity report hours is $r = 0.11$, n.s. The correlation with code insertions/deletions/changes is $r = -0.08$, n.s. Based on these results we reject H2.

To test H3, we converted the survey responses (perceived communication) and minute counts (observed communication) to z-values per participant, then computed the integer value of $(z_{\text{perceived}} - z_{\text{oberved}}) * 10$. The distribution is shown in Fig. 2. The mode is zero, indicating that the largest number of cases had no difference between the perceived and observed values. The distribution is not normal (W = 0.81, 93 d.f., $p < 0.001$). It has left skewness of 1.49, indicating that participants over-report communication with others more than they under-report.

We correlated these values, summed over the study period for all pairs of participants, with the edge weights in the one-mode (participant) projection of the two-mode network linking participants with projects. The correlation with the standardized perceived-observed difference was $r = 0.40$, $p < 0.01$. We also correlated the perceived/observed difference with the number of times one participant reported being paired with another. Here the correlation was $r = 0.36$, $p < 0.01$. These results support H3. We note that the project and partner measures were themselves correlated, $r = 0.55$, $p < 0.01$.

Discussion

This research revisited the network accuracy studies using a unique dataset. We collected data over an extended period of 144 weeks in a naturalistic organizational setting. This resolves the methodological criticisms noted above that results of the accuracy studies may have been influenced by unusual contexts where communication is routinely logged or by the sampling methods used to do manual observation in previous research. The extended observation period also allows us to assess the extent to which perceived/observed correlations vary over time under similar observational circumstances.

Our first hypothesis predicted that communication, as reported by participants, is a valid predictor of observable communication between those participants. Two approaches exist in the literature for testing this hypothesis, one analyzing the association between perceived and observed measurements at the dyadic level, and another looking at the structural similarity of networks derived from the dyadic data. Regarding the former, QAP correlations between weekly adjacency matrices recording perceived and observed communication were significant for only about 70 % of the weeks. Correlations averaged $r = 0.25$ for all weeks, and $r = 0.32$ for significant weeks. These correlations are somewhat lower than those reported in the accuracy studies, where correlations averaged $r = 0.45$ across studies. They also fall within the ranges reported by other studies reported above looking at behavior recall outside network contexts.

Importantly, this study also shows that there is considerable variation in perceived/observed correlations across the weeks studied, with a standard deviation of 0.32. Since all the correlations were based on the same overall set of participants and the same perceived/observed measures, this variation cannot be due to measurement techniques or changes in participants. This means that the results in the accuracy studies might also have differed from what was observed, had the data been collected during different time periods.

At the structural level, we observed significant relationships between perceived and observed networks in terms of correlation transitivity ($r = 0.29$), and edge weight marginals ($r_{\text{inbound}} = 0.26$, $r_{\text{outbound}} = 0.41$). Values for correlation transitivity were significantly lower for perceived networks ($M = 0.56$) than for observed networks ($M = 0.69$). Overall then, we find that almost one-third of the weeks do not have significant dyadic correlations, and for the weeks that do, perceived and observable measures share about 26 % of their variance. The structural level measures show slightly lower values with shared variance of up to 17 %, but even at this level we cannot conclude that reports by participants of perceived communication accurately predict communication that can be observed.

Our tests of H2 fail to replicate the results of Corman and Bradford, 1993. Results showed that the workload of the participants, operationalized as hours logged in the activity reporting system or code repository activity, was not significant predictor of the perceived/observed correlations in a given week. This is at least partly because employees did not always log hours in the metrics reporting system, and SF work involved much more than writing code (i.e., researching solutions, planning, code reviews, etc.). The tests of H3 do replicate the findings of Author, showing that the more one participant is formally connected with another, either through co-work on projects or partnering relationships, the more they tend to overestimate communication with that other. Finally, we note that there is a general bias toward overreporting. The error distribution shown in Fig. 2 is showing that participants over-reported more than they under-reported. This is consistent with results reported by Kobayashi and Boase (2012) in the context of mobile phone use.

Limitations

Four limitations of this study deserve mention. First, our method for detecting communication from audio recordings only looked at the presence of a common voice signal on two recordings. We may have misclassified some cases as communication where, for example, one person was talking on the phone in the vicinity of another person being recorded. Given the number of recordings, it was not practical to verify that these cases marked an actual conversation, but our coders indicated that such cases were rare. Second, we did not replicate all the structural equivalence methods used by the critics of the accuracy studies because of differences in data.

A third limitation has to do with the generalizability of our findings, given that they are based on a single organization. We believe the diversity of SF projects, the ubiquity of the observations, the extended period over which they were gathered, and the difficulty of reproducing this effort over many organizations makes the generalizability concerns tolerable. Our findings differ from those of previous studies, but not radically so, easing concerns the SF is an outlier organization.

Finally, we did not have access to data on all possible sources of communication load for the test of H2. For example, we do not have participants' email messages or telephone logs. However, SF was located in one large, open office setting so participants could easily communicate without using email or telephone, so we do not believe these are likely large sources of load.

Implications

There are three take-aways from this study. First, correlations reported here are somewhat lower than those reported in the accuracy study. At the dyadic level, perceived communication accounts for 6.25 % (all weeks) to 15.36 % (QAP-significant weeks) of the variance in observable communication. At the structural level, perceived structures account for up to 17 % of the variance in observed structures, depending on the measure used. This value is lower than many structure-level correlations reported by accuracy study critics; however, we note that the way our data were collected and compiled precluded exact replication of some of the critics' methods.

Second, there is considerable variation in dyadic correlations over the many weeks we studied. For about 42 % of the weeks we studied, correlations did not exceed the QAP threshold for significance. Those that were significant varied over a range of up to 20 % of variance explained. Unlike the accuracy studies, we studied the same organization with the same members (except for normal turnover) using the same methods, so these differences must be due to something other than methodological artifacts. An intriguing possibility is that variation in accuracy is the outcome of social forgetting that allows the organization to be more adaptive.

Third, some of these differences are due to situational factors. A heavier workload among participants included in a given week's analysis is not associated with a lower perceived-observed correlation for the same week. However, participants do tend to overreport communication with others with whom they share assignments to projects or are frequently partnered for tasks. In other words, self- reports are biased toward formal structural relationships.

When all is said and done, the conclusion of Bernard et al. (1984), that perceived network data bears *no* useful resemblance to communication that occurs, may be going too far. Our results show significant correlations for about half of the weeks. Perceived communication explains up to about one-sixth of the variance in observed interaction, and in many research contexts (for example a social psychology study) that would be considered a useful amount. In many cases, participants accurately report their communication or differ from what can be observed by small amounts, and there are sometimes large correlations between perceived and observed communication at the network level. However, the correlations vary over a wide range from week to week, and that situational factors like organizational structures bias self-reports.

The "no useful resemblance" conclusion of Bernard and colleagues assumes that the lack of high correlations means perceived network measures are simply flawed indicators of corresponding observed network measures. But if that were true, we would expect more consistent correlations across studies as well as less variation from week to week in the longitudinal results presented here. On the other hand, we *would* expect variation in correlations if perceptions were one of the influences—but not the only influence—on observable behavior.

This points to the importance of developing and testing theory that more fully explains the relationship between perceived and observable communication. One example is network reticulation theory (NRT; Author; Fan et al., 2020). Using a structurational approach, NRT argues that perceived networks are structures in a latent domain of social relations. When we ask participants about their perceived network relationships with others, this is what we are measuring. Perceived networks are activated by focused activity in an organizational setting, aimed at accomplishing some task, goal, or requirement, and leading to behavior in an observable domain of social interaction. In any activity demand, it might be that an existing strong relationship is activated, and this is probably the norm. However, circumstances (time pressures, availability of members, etc.) might dictate that a weaker relationship be activated or that a new one be formed, leading to a mismatch between perceived and observed behavior at that time.

Observable behavior in the domain of social interaction in turn influences the perceived relationship in the domain of social structure, strengthening or weakening it. This production and reproduction of perceived networks in observable interaction over time is a plausible reason that a given instance of observable behavior might not correspond to a perceived relationship: Circumstances of activation might have prevented it, or the perceived relationship might have not yet solidified. A complex system drives the communication choices of organization members, and perceived and observable networks are distinct yet related elements of that system. Further empirical research is needed to test this proposition.

Acknowledgements

This research was supported in part by a grant from the U.S. National Science Foundation's Science of Organizations program (award 1632707). Per NSF requirements, data from this project will be made available through the ICPSR repository at https://doi.org/10.3886/E130041V1.

The authors thank Steve Borgatti for assistance with automating our QAP analysis on UCINET, and Kevin Dooley and Steve Gordon for assistance with organizational archaeology.

Appendix A

To detect observable communication from audio recordings we used cross channel speech signal analysis to detect communication between speakers. The basic idea behind this approach is that, if two individuals are speaking, their microphones will pick up each-other's speech and cross correlation will be high. The following steps are used for cross channel speech analysis. As a pre-processing step we normalized the data by the mean to remove DC offset (caused by the analogue parts of the system that add a DC current to the audio signal), that causes significant interference with the audio signal, especially during signal processing. We investigated preliminary conversation detection performance on the SF data by using a two-stage approach. The first stage identified continuous segments of speech using an energy and spectral based detector; in the second stage, we use pair-wise cross-correlation between one speaker's channel and the remaining channels to detect with whom that person was speaking. We computed the short-time speech energy and spectral centroid for every 15 s frame and estimated thresholds to detect speech from the two features. Speech portions were detected using the two thresholds and nonspeech portions were removed. Next, we computed the covariance matrix between energy of the speech segments from both microphones in a dyad. The covariance matrix represents a proxy for the frequency of interactions between any two individuals. Two sets of thresholds were estimated based on the diagonal elements of the matrix, (a) Th_1, to determine if communication occurred (0 or 1, 2, 3) and (b) Th_2, to determine the direction of communication (1, 2 or 3).

We validated the detections by comparing them to human coder classifications of the audio recordings as indicating network connections. We extracted 10-minute audio segments from a dyad from random working days. First, we determined the total number of segments required to assess validity. Based on this we extracted that number of segments through random sampling from the audio corpus. External raters then coded the 15 s segments regarding whether there was talk or silence in the segment and who was talking to whom. The specific classifications they could make were: Silence/noise (0), Employee 1 speaking (1), Employee 2 speaking (2), Both employees speaking (3). We determined the minimum number of audio segments required to assess validity using the confidence interval equation,

$$N \geq \frac{\widehat{p}(1-\widehat{p})}{\in^2} \tag{1}$$

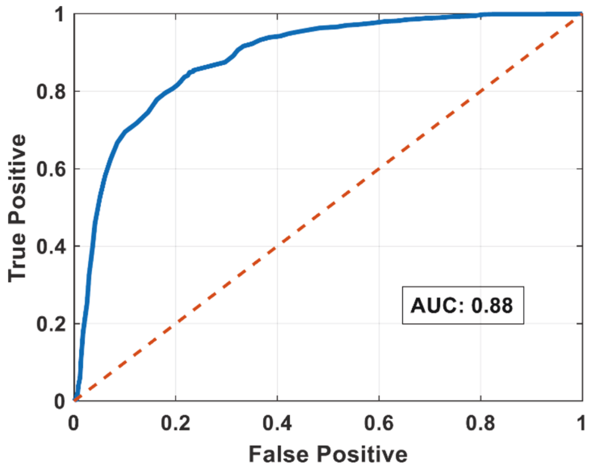


Fig. A1. Receiver operating characteristics curve for communication detection; Area under curve (AUC) = 0.88.

Table A1
Confusion matrix for the best detection model. Each element is shown in terms of number of 15 s segments.

	Coder				
	Class	0	1	2	3
Tool	0	1390	183	227	74
	1	51	201	32	105
	2	70	14	309	98
	3	16	42	50	138

where N is the minimum number of samples, $\widehat{p}$ is the estimated population proportion and $\in$ is the margin of error. With an error margin (variance per sample) of 5% and a $\widehat{p}$ of 0.8, the minimum number of samples required is 64. In our analyses, a total of 75 ten minutes audio segments from random working days and between random dyads were used for communication validation. As Fig. A1 indicates, there was 88 % agreement between the coders and the automated detection (see next section for more details).

In the pair-wise communication detection, the four main classes were, "*Silence/noise*" (0), "*Employee 1 speaking*" (1), "*Employee 2 speaking*" (2), and "*Both employees speaking*" (3). The receiver operating characteristics (ROC) curve (see Fig. A1) was used to illustrate the communication detection accuracy (0 or 1, 2, 3). The ROC curve was constructed by varying the threshold Th_1, and the optimum value of Th_1 was determined. Threshold Th_2 was determined after constructing confusion matrices for various Th_2 values. The threshold parameters for the best model were $Th_1 = 2.53e^{-5}$ and $Th_2 = 2.02e^{-5}$. We have shown the confusion matrix of the best detection model (80 % training, 20 % testing) in Table A1.

Our method produced a communication detection rate (AUC: 0.88), and on reviewing the results, we noticed that most of the false positives resulted because of the presence of other employees. Thus, in case of a communication scenario with more than two employees, the correlation weights will be high for any dyad with the speaker in it, while the correlation weights between other employees will be relatively low. For any focal individual, the correlation weights between that individual will be high with anyone they address, while those between other speakers who might be detected in the background is lower.

Short-time speech energy and spectral centroid

We denote $x_i(n)n = 1, ..., W_L$ as the sequence of audio samples of the i-th frame, where W_L is the length of the frame. Short-time energy distinguishes voiced speech from unvoiced speech and evaluates the amplitude variation and power of the signal for each frame. It is calculated as,

$$E(i) = \sum_{n=1}^{W_L} |x_i(n)|^2 \tag{2}$$

Let $X_i(k)k = 1, ..., W_{FL}$, be the magnitude of the DFT coefficients of the i-th audio frame. Spectral centroid is a measure of the center of gravity of the spectrum of the signal frame. A higher value of spectral centroid corresponds to a brighter sound. It is calculated as:

$$C = \frac{\sum_{k=1}^{W_{FL}} kX_i(k)}{\sum_{k=1}^{W_{FL}} X_i(k)} \tag{3}$$

References

Bernard, H.R., Killworth, P.D., Sailer, L., 1982. Informant accuracy in social-network data V. An experimental attempt to predict actual communication from recall data. Soc. Sci. Res. 11, 30–66. https://doi.org/10.1016/0049-089X(82)90006-0.

Boase, J., Ling, R., 2013. Measuring mobile phone use: self-report versus log data. J. Comput. Commun. 18, 508–519. https://doi.org/10.1111/jcc4.12021.

Corman, S.R., Scott, C.R., 1994a. Perceived Networks, Activity Foci, and Observable Communication in Social Collectives. Commun. Theory 4, 171–190. https://doi.org/10.1111/j.1468-2885.1994.tb00089.x.

Corman, S.R., Scott, C.R., 1994b. A synchronous digital signal processing method for detecting face-to-face organizational communication behavior. Soc. Networks 16, 163–179. https://doi.org/10.1016/0378-8733(94)90003-5.

Bernard, H.R., Killworth, P., 1977. Informant accuracy in social network data II. Hum. Commun. Res. 4, 3–18. https://doi.org/10.1111/j.1468-2958.1977.tb00591.x.

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. Soc. Netw. 2, 191–218. https://doi.org/10.1016/0378-8733(79)90014-5.

Bernard, H.R., Killworth, P.D., Kronenfeld, D., Sailer, L., 1984. The problem of informant accuracy: the validity of retrospective data. Annu. Rev. Anthropol. 13, 495–517. https://doi.org/10.1146/annurev.anthro.13.1.495.

Brewer, D.D., 2000. Forgetting in the recall-based elicitation of personal and social networks. Soc. Netw. 22 (1), 29–43. https://doi.org/10.1016/S0378-8733(99)00017-9.

Burt, R.S., Bittner, W., 1981. A note on inferences regarding network subgroups. Soc. Netw. 3, 71–88. https://doi.org/10.1016/0378-8733(81)90006-X.

Coombs, C.H., 1958. An applicaiton of a nonmetric model for multidimentional analysis of similarities. Psychol. Rep. 4, 511. https://doi.org/10.2466/PR0.4..511-518.

Corman, S.R., 1990. A model of perceived communication in collective networks. Hum. Commun. Res. 16, 582–602. https://doi.org/10.1111/j.1468-2958.1990.tb00223.x.

Corman, S.R., Bradford, L., 1993. Situational effects on the accuracy of self-reported organizational communication behavior. Communic. Res. 20, 822–840.

Dekker, D., Krackhardt, D., Snijders, T.A.B., 2017. Transitivity Correlation: Measuring Network Transitivity as Comparative Quantity. arXiv Prepr. arXiv1708.00656.

Fan, C., Shen, J., Mostafavi, A., Hu, X., 2020. Characterizing reticulation in online social networks during disasters. Appl. Netw. Sci. 5, 10–29. https://doi.org/10.1007/s41109-020-00271-5.

Freeman, L., Romney, A., 1987. Words, deeds and social structure: a preliminary study of the reliability of informants. Hum. Organ. 46, 330–334. https://doi.org/10.17730/humo.46.4.u122402864140315.

Freeman, L.C., Romney, A.K., Freeman, S.C., 1987. Cognitive structure and informant accuracy. Am. Anthropol. 89, 310–325. https://doi.org/10.1525/aa.1987.89.2.02a00020.

Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. 1, 77–89. https://doi.org/10.1080/19312450709336664.

Holland, P.W., Leinhardt, S., 1975. Local structure in social networks.". In: Heise, D. (Ed.), Sociological Methodology. Jossey-Bass, San Francisco.

Huisman, M., Steglich, C., 2008. Treatment of non-response in longitudinal network studies. Soc. Netw. 30 (4), 297–308. https://doi.org/10.1016/j.socnet.2008.04.004.

Johnson, J.C., Miller, M.L., 1986. Behavioral and cognitive data: A note on the multiplexity of network subgroups. Soc. Netw. 8 (1), 65–77. https://doi.org/10.1016/S0378-8733(86)80015-6.

Kashy, D.A., Kenny, D.A., 1990. Do you know whom you were with a week ago friday? A re-analysis of the bernard, killworth, and sailer studies. Soc. Psychol. Q. 53, 55. https://doi.org/10.2307/2786869.

Killworth, P., Bernard, H.R., 1976. Informant accuracy in social network data. Hum. Organ. 35, 269–286. https://doi.org/10.17730/humo.35.3.10215j2m359266n2.

Killworth, P.D., Bernard, H.R., 1979. Informant accuracy in social network data III: a comparison of triadic structure in behavioral and cognitive data. Soc. Netw. 2, 19–46. https://doi.org/10.1016/0378-8733(79)90009-1.

Killworth, P.D., McCarty, C., Bernard, H.R., House, M., 2006. The accuracy of small world chains in social networks. Soc. Netw. 28, 85–96. https://doi.org/10.1016/j.socnet.2005.06.001.

Kimball Romney, A., Weller, S.C., 1984. Predicting informant accuracy from patterns of recall among individuals. Soc. Netw. 6, 59–77. https://doi.org/10.1016/0378-8733(84)90004-2.

Kobayashi, T., Boase, J., 2012. No such effect? The implications of measurement error in self-report measures of mobile communication use. Commun. Methods Meas. 6, 126–143. https://doi.org/10.1080/19312458.2012.679243.

Krackhardt, D., 1988. Predicting with networks: nonparametric multiple regression analysis of dyadic data. Soc. Netw. 10, 359–381. https://doi.org/10.1016/0378-8733(88)90004-4.

Menon, G., 1993. The effects of accessibility of information in memory on judgments of behavioral frequencies. J. Consum. Res. 20, 431–440.

Neyman, J., 1937. Outline of a theory of statistical estimation based on the classical theory of probability. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 236, 333–380. https://doi.org/10.1098/rsta.1937.0005.

Powers, D.M.W., 2011. Evaluation: from precision, recall and F-Measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2, 37–63.

Romney, A.K., Faust, K., 1982. Predicting the structure of a communications network from recalled data. Soc. Netw. 4, 285–304. https://doi.org/10.1016/0378-8733(82)90015-6.

Schwartz, N., 1990. Assessing frequency reports of mundane behaviors. In: Hendrick, C., S, C.M (Eds.), Research Methods in Personality and Social Psychology. Sage Publications, Newbury Park, CA, pp. 98–119.

Shrout, P.E., Rodgers, J.L., 2018. Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. Annu. Rev. Psychol. https://doi.org/10.1146/annurev-psych-122216-011845.

Singh, V.K., Jain, A., 2017. Toward harmonizing self-reported and logged social data for understanding human behavior. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI' 17. ACM Press, New York, New York, USA, pp. 2233–2238. https://doi.org/10.1145/3025453.3025856.

Webster, C.M., 1992. Seeing is believing: the use of cognitive measures in determining group structure. In: 12th International Sunbelt Social Networks Conference. San Diego, CA.