Testing Influence of Network Structure on Team Performance Using STERGM-Based Controls

Brennan Antone[1(✉)], Aryaman Gupta[1], Suzanne Bell[2], Leslie DeChurch[1], and Noshir Contractor[1]

[1] Northwestern University, Evanston, IL 60201, USA
brennanantone2017@u.northwestenr.edu
[2] DePaul University, Chicago, IL 60614, USA

Abstract. We demonstrate an approach to perform significance testing on the association between two different network-level properties, based on the observation of multiple networks over time. This approach may be applied, for instance, to evaluate how patterns of social relationships within teams are associated with team performance on different tasks. We apply this approach to understand the team processes of crews in long-duration space exploration analogs. Using data collected from crews in NASA analogs, we identify how interpersonal network patterns among crew members relate to performance on various tasks. In our significance testing, we control for complex interdependencies between network ties: structural patterns, such as reciprocity, and temporal patterns in how ties tend to form or dissolve over time. To accomplish this, Separable Temporal Exponential Random Graph Models (STERGMs) are used as a parametric approach for sampling from the null distribution, in order to calculate p-values.

Keywords: Network properties · Team performance · Separable temporal exponential random graph models

1 Introduction

Across many areas of network science, research often asks questions about how different network-level properties or outcomes are related. For instance, when studying networks between team members, researchers may examine whether properties of the whole team's network, such as density or centralization, impact the performance of that team. In this case, each network may have one score for density and one score for team performance, and through the observation of multiple networks, a correlation between density and team performance can be computed. However, it is important to assess the potential spuriousness of these correlations, especially when working with a small sample of networks.

When correlations involve *network statistics*, functions that return a single score computed from a network, proper significance testing may be challenging.

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2019, SCI 882, pp. 1018–1030, 2020.
https://doi.org/10.1007/978-3-030-36683-4_81

The ties and node attributes that determine the value of the network statistic often are not statistically independent of one another. Complex interdependencies may exist between ties and node attributes within the same network: for instance, ties may tend to be reciprocated, or ties may be more likely between nodes of the same gender. We will refer to this type of interdependency as *structural patterns* in how ties form. Additionally, in the event that researchers have collected data at multiple points in time, there may be complex interdependencies between ties in repeat observations of the same nodes. For instance, if a tie exists between individuals at one point in time, it is natural to expect that tie to be more likely to exist the next time these individuals are observed. We will refer to this type of interdependency as *temporal patterns* in how ties form.

We argue that, when the value of network statistics being observed may be influenced by structural patterns or temporal patterns in how network form, significance tests of correlations involving these network statistics need to control for these patterns. An existing modeling approach, Separable Temporal Exponential Random Graph Models (STERGM) provides a way to identify both structural and temporal patterns [7]. We demonstrate how, when computing p-values for correlations involving network statistics, STERGMs can be used as a parametric approach for sampling from the null distribution.

We will apply this approach to understand team processes in the crews of long-duration space exploration (LDSE) missions, linking different patterns of social relations between crew members to measures of crew performance on various tasks. This form of significance testing is particularly beneficial to data collected in LDSE analogs such as those operated by NASA (e.g., Human Exploration Research Analog [HERA]), in which relatively few crews may be observed, but at multiple points in time. We demonstrate how crew performance on different tasks benefits from different types of network structure.

2 Motivation

To identify network influences on team performance, our goal is to quantify the possibility that observed correlations may be spurious. We frame our analysis around the null hypothesis that there is, in fact, no correlation between any two variables being tested. In this case, these variables will be some network statistic computed from a team's network and team performance. Based on the data we collected, we can compute an *observed correlation coefficient*, and an *empirical p-value*. The empirical p-value reflects the probability that we would find a correlation coefficient equal to or more extreme than the observed correlation coefficient in the event that there was no correlation between the network statistic and team performance. We will consider what issues structural and temporal patterns may introduce for such an analysis.

2.1 Influence of Structural Patterns

To understand the impact of structural patterns on correlation coefficients, consider a study of four-person teams. If we were to examine the network statistic of

closeness centralization [4], for a directed network there are 27 possible scores a four-person team can have for closeness centralization. Closeness centralization is a deterministic function of the 12 ties in the team. The likelihood of observing each of the 27 levels of closeness centralization may vary depending on structural patterns present, such as reciprocity, closure, or homophily. Under the influence of different structural patterns, different values of network statistics, and therefore correlation coefficients, may be more or less likely to occur by chance alone. Considering this, knowledge about structural patterns should inform how "surprising" it would be to observe a given correlation coefficient in the event that our null hypothesis is true.

2.2 Influence of Temporal Patterns

Further complications may be introduced if a researcher wishes to incorporate multiple observations of a team's performance over time. Repeat observations may contain new information to help inform conclusions about team performance. To leverage such observations, we would want to control for non-independence (ex. autocorrelations) between the network statistics in repeat observations of a team. This can be accomplished by modeling temporal dependencies between a tie's current value, that tie's past or future values, and other ties' past or future values. These may be simple trends, like the tendency of a tie to continue existing over time, or more complex trends, like the tendency of ties to form if doing so would complete a transitive triad. If there is not a way to control for these types of patterns, using repeated observations of a team would be problematic. Using multiple observations of a team may be critical in research contexts where obtaining data from additional teams may be costly or impossible, but teams are able to be studied for an extended period of time.

2.3 Existing Approaches

The development of null models for significance testing has a long history in network science, because interdependencies between ties must be controlled for when performing significance testing. Two fundamental approaches for significance testing when dealing with such interdependencies are nonparametric approaches and parametric approaches for sampling from the null distribution. *Nonparametric approaches* often rely on a type of permutation test, in which observations are shuffled in some systematic way. For example, Quadratic Assignment Procedure (QAP) and its extension multiple regression QAP (MRQAP) are often used to test correlations between the presence of ties in two or more networks by performing random relabeling of nodes in each permutation while keeping network structure constant [5,6]. Alternatively, approaches based on network rewiring are often used to permute the location of ties or events in a network while maintaining properties of that network's degree distribution [12,14].

Parametric approaches entail the estimation of a model to sample from the null distribution, the distribution of the test statistic (ex. correlation coefficient) that would be observed in the event that the null hypothesis was true. Whereas

nonparametric approaches all either maintain network structure or specify exact rules for how network structure should be permuted, a parametric approach can estimate, based on data, the types of complex interdependencies that exist and perform appropriate permutations to control for them.

We propose a parametric approach for significance tests involving network-level statistics, in which both structural patterns and temporal patterns are modeled using STERGM. This approach will offer the distinct benefit of allowing observations of networks at multiple points in time to be used, by modeling how networks change between observations when sampling from the null distribution.

3 Approach for Testing the Influence of Network Structure on Team-Level Performance

Let us assume we have observed networks between team members, including different node attributes or other exogenous attributes that may affect network formation. Some of these networks may be collected from the same team at multiple points in time. We will refer to an ordered set of networks we collect from the same team as a *temporal path of networks*. We also assume that we have measured some *performance metric*, which assigns a single score to each network. Finally, we assume we have chosen a *network statistic*, a deterministic function of the ties and node attributes in a network, that we are interested in correlating with our performance metric. Our goal will be to assess the probability our sample might produce a correlation as extreme as the one observed if there was truly no correlation between our network statistic and performance metric.

We begin by calculating the correlation between the network statistic for each network and the corresponding performance metric from our empirical data. This produces the *empirically observed correlation coefficient*. While any form of correlation could be used, we suggest that due to the discrete or non-normally distributed nature of many network statistics it would often be appropriate to use a Spearman rank correlation coefficient [17]. Our approach will aim to estimate the probability that we would observe a correlation coefficient that extreme, if there was truly no correlation between our network statistic and performance metric in the full population. This is the empirical p-value. To conduct such a test, we want to control for structural and temporal patterns shaping how networks form. To accomplish this, we must define a version of the *null model* that accounts for each of these trends that may occur in the observed networks.

We propose using Separable Temporal Exponential Random Graph models (STERGM) as a flexible framework to construct our null models [7]. STERGMs describe how networks are likely to change over time by defining a joint probability distribution for the presence of ties in a series of repeat observations of networks. This distribution is defined by two sets of assumed sufficient statistics: A vector of *formation statistics* $\mathbf{g}^{-}()$ and their corresponding weights $\boldsymbol{\theta}^{-}$ describe how likely it is that a subset of ties $\mathbf{Y}^{-}$ that did not exist in a previous network $\mathbf{Y}^{\mathbf{t}}$ will form. A vector of *persistence statistics* $\mathbf{g}^{+}()$ and their corresponding weights $\boldsymbol{\theta}^{+}$ describe how likely it is that the subset of ties $\mathbf{Y}^{+}$

that existed in a previous network $\mathbf{Y^t}$ will continue to exist. In both cases, the assumed sufficient statistics are a function of both the ties being predicted and some dyadic or node covariates $\mathbf{X}$. The joint probability of each subset of ties, for a network at a single point in time, is expressed as:

$$P(\mathbf{Y}^- = \mathbf{y}^-) = \frac{e^{\boldsymbol{\theta}^- \mathbf{g}^-(\mathbf{y}^-, \mathbf{X})}}{\sum_{\mathbf{i} \in \mathbf{Y}^-} e^{\boldsymbol{\theta}^- \mathbf{g}^-(\mathbf{i}, \mathbf{X})}} \tag{1}$$

$$P(\mathbf{Y}^+ = \mathbf{y}^+) = \frac{e^{\boldsymbol{\theta}^+ \mathbf{g}^+(\mathbf{y}^+, \mathbf{X})}}{\sum_{\mathbf{i} \in \mathbf{Y}^+} e^{\boldsymbol{\theta}^+ \mathbf{g}^+(\mathbf{i}, \mathbf{X})}} \tag{2}$$

STERGMs are estimated using conditional maximum likelihood estimation as described in Krivitsky & Handcock 2014, in our case applying this technique to estimate a single model that jointly captures trends that occur amongst all of the temporal paths of networks we observed. As part of this estimation, as with any STERGM model, model convergence and goodness of fit should be assessed in order to make sure that the parameter estimation was successful and that the model replicates trends observed in the empirical data.

In comparison to TERGMs or ERGMs, STERGMs offer the benefit of representing complex temporal patterns in either the formation or persistence of ties between observations. Thus, when using multiple observations from the same networks over time, this provides an explicit mechanism for controlling for temporal dependencies between them when sampling from the null distribution.

STERGMs fit to our data are used to sample correlation coefficients from the null distribution. We simulate random networks according to the STERGM using Markov Chain Monte Carlo (MCMC) sampling. To obtain a simulated correlation coefficient, we take each different team in our dataset and simulate that team's temporal path of networks based on the node attributes and exogenous factor values for that team. We then calculate a correlation between the network statistics for all of the simulated networks and the performance metric for each network that we observed in our empirical data. By repeating this, we obtain a sample that approximates the null distribution, the distribution of correlation coefficients we would obtain based on our null model.

We then compare the empirically observed correlation coefficient to our samples from the null distribution. If we let n denote the total sample size of simulated correlation coefficients and k as the count of simulated correlation coefficients that are at least as extreme as our observed correlation coefficient (greater than or equal to if positive, less than or equal to if negative), then we estimate the p-value as the ratio k/n.

4 Application: Relational Indicators of Crew Success in Long-Duration Space Exploration

We will examine how different patterns of social relations between crew members of long-duration space exploration missions are associated with crew performance. Future lunar and Mars missions will entail extended trips, in which a

small crew must work together more effectively and autonomously, since communication delays will grow as the crew travels away from the earth. Thus, effective team processes will be critical to team success [15]. By understanding the effects of social networks on crew performance, space agencies will better able to staff and support crews on these missions by examining their interpersonal relations.

A challenge in research about long-duration space exploration (LDSE) is the limited ability of relevant data. One environment for collecting data is LDSE-analogs, in which participants may complete tasks typical of LDSE while living in an isolated environment for an extended time [9]. These analogs allow researchers to collect high-quality data from only a small number of crews over an extended time. Because of this, there is a need for analysis capable of leveraging repeated observations of a team to test what factors impact crew performance [2].

4.1 Measures

Research Setting. Data was collected from the Human Exploration Research Analog (HERA), an extended simulation of space exploration that is operated by NASA at the Johnson Space Center in Houston, Texas. Participants in HERA missions completed tasks over the course of a 30 or 45 day mission that simulated space exploration, remaining confined in the HERA capsule for the entire duration. Over the course of missions, participants experienced long shifts, sleep deprivation, communication delay with ground control, and emergency simulations designed by NASA to provide a realistic simulation of space exploration.

Respondents. Eight four-person crews completed analog space missions between January 2016 and June 2018. Four crews completed 30 day missions, and four completed 45 day missions. Each crew had a designated commander, a flight engineer, and two mission specialists. Of the 32 respondents, 59.4% were female, the average age was 38.0 years (s.d. = 7.98), and 34.4% had military experience. When asked about their race/ethnicity, 24 respondents selected Caucasian non-Hispanic, two selected Caucasian Hispanic, two selected East Asian, one selected South/Southeast Asian, one selected African American, and two selected "Other".

Performance Dimensions. Team performance measures the degree to which a team accomplishes its goals. Four dimensions of performance are summarized by McGrath's Task Circumplex [13]. We used measures derived from four tasks reported in Larson et al., 2019 [10]. The *Generate* task required the crew to develop new ideas. The *Choose* task required them to solve a survival scenario with a known solution. The *Negotiate* task required them to resolve an ethical dilemma incorporating multiple conflicting viewpoints. The *Execute* task was a simulation in which a pilot and co-pilot use a joystick to fly a transit vehicle to collection sites, while the other two crew members use virtual reality goggles to complete Extra Vehicular Activity exploring an asteroid's surface. For the four 30-day missions, these tasks were administered three times, on mission days 10, 15, and 29.

For the four 45-day missions, the tasks were completed four times, on days 13, 18, 27, and 41. This produced a total of 28 observations of team performance.

Social Relations. Social networks were elicited from the crew via sociometric surveys. We included measures of four relational networks to capture a long-standing distinction in the small group literature on task and social needs. *Task affect* and *task hindrance* capture positive and negative working relationships among crew members. Task affect was measured with the prompt: "With whom do you enjoy working?" Task hindrance was elicited with the prompt: "Who makes tasks difficult to complete?" In addition to assessing manifest social relations, we also included two networks capturing behavioral and motivational aspects of teams: *leadership* and *followership*. Leadership was elicited by asking "To whom do you provide leadership?" Followership relations were assessed by asking: "Who do you rely on for leadership?" These prompts yielded four directed networks, each examined in relation to performance. Performance scores from each task session were matched with the network survey most closely proceeding the task. For the 30-day missions, social networks and performance, respectively were measured on the following pairs of days: days 9 and 10, 14 and 15, and 27 and 29. For the 45-day missions, networks and performance events, respectively, were measured on these pairs of days: 11 and 13, 15 and 18, 26 and 27, and 39 and 41.

Network Statistics. While network *density*, the ratio of observed to possible ties, may influence performance directly, where ties are located relative to one another may also impact performance (Fig. 1). Network theories of teams posit the degree of closure, centralization, and subgrouping among team members are important reflections of the quality of teamwork and the team's capacity to perform [3]. We selected six network statistics based on these three categories, in addition density, to test their impact on crew performance.

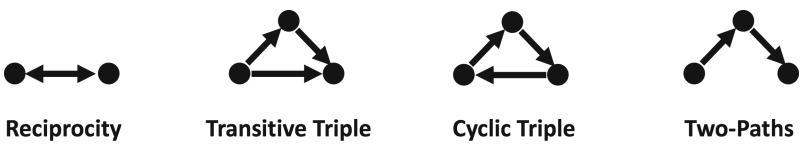

Fig. 1. Basic network structures used in defining network statistics

For the closure category, we measured *normalized transitivity* and *normalized cyclicality*. Normalized transitivity controls for density effects, computing transitivity of a graph as the number of transitive triples divided by the average number of transitive triples in a random graph of the same density. The same approach was used to define a statistic for normalized cyclicality, normalizing by the expected number of cyclic triples.

To examine centralization, we include *closeness centralization*, relative discrepancies in closeness, as defined in Freeman 1978, between team members. Closeness centrality ranks team members based on the proportion of the shortest paths between all team members on which they lie. A team in which each member had a similar closeness centrality would have a lower value of closeness centralization, whereas a team with big differences in closeness centralization would have a high value. Given that our crews include a team member assigned to the role of commander, we also examined centralization using the *relative indegree of commander*, measuring the proportion of all ties directed towards the commander, and the *relative outdegree of commander*, measuring the proportion of all ties directed from the commander towards others.

To examine subgrouping in four-person networks, we consider the amount of two-paths, chains of ties spanning between three crew members, as a way of measuring a tendency against subgrouping. We include a statistic for *normalized two-paths*, using the same approach to normalizing the count of two-paths based on density as used for transitivity and cyclicality.

Controls for Structural and Temporal Patterns. Because teams are observed at multiple points in times, null models need to control for repeated observations of the same team. This is accomplished by including corresponding sufficient statistics in the STERGMs. The first temporal pattern we control for is tie formation, the likelihood that a new tie will form where a tie had not previously existed, by including a *tie likelihood* term in the formation model that counts the number of ties in the network. Similarly, we also control for tie persistence, the likelihood that a tie that has previously existed will continue to exist, by including a tie likelihood term in the persistence model.

Next, we considered the potential effects of extended isolation, in which a crew is forced to work and live together while working long shifts. We included terms for the effects of *elapsed time in isolation* on the likelihood of new ties to form and for effects of elapsed time in isolation on the likelihood of existing ties to persist. We also included a *time between observations* term that examines the effect of the time, in days, since the network data were last collected. This controls for the fact that our data was not collected in uniformly spaced intervals.

Another well documented structural pattern is *reciprocity*. We include measures for the count of reciprocated ties in our STERGMs, to control for tendencies of new ties to be more or less likely to form reciprocated pairs, as well as for existing ties in reciprocated pairs to be more or less likely to persist. Finally, we control for potential homophily effects in our models for formation and persistence terms for *race homophily* and *military experience homophily*.

4.2 Analysis

To develop a null model for our analysis, we estimated one STERGM for each of the four ties, using the temporal networks collected from all teams. STERGMs were fit using Conditional Maximum Likelihood Estimation [7], as implemented

in the tergm package developed for R [8]. To measure the association between a network statistics and performance metric, we used Spearman rank correlation coefficients [17]. P-values were computed for these correlation coefficients using our approach for sampling from the coefficients' null distribution. A total of 250 simulated values of correlation coefficients were utilized.

4.3 Results

Descriptive Results. Intercorrelations between performance scores across the four task dimensions, as well as intercorrelations for the presence of ties in the four networks, are reported in Table 1. In particular, we note that the Spearman rank correlation coefficient between any two of the performance dimensions ranged between −0.62 and 0.38. Because they are not perfectly correlated, it is critical we separately analyze the associations between network structure and each dimension of team performance. The 28 task affect networks had an average density of 0.83 (s.d. = 0.13). Task hindrance networks had an average density of 0.23 (s.d. = 0.16), leadership networks had an average density of 0.79 (s.d. = 0.17), and followership networks had an average density of 0.63 (s.d. = 0.22).

Table 1. Intercorrelations for team performance and network ties

Performance Measure Intercorrelations					Network Tie Intercorrelations				
	Generate	Choose	Negotiate	Execute		Task Affect	Task Hindrance	Leadership	Followership
Generate	-	0.15	-0.23	-0.11	Task Affect	-	-0.49	0.11	0.39
Choose		-	-0.28	-0.62	Task Hindrance		-	0.06	-0.28
Negotiate			-	0.38	Leadership			-	0.17
Execute				-	Followership				-

Table 2. STERGM results for each social relation

	Task Affect		Task Hindrance		Leadership		Followership	
	Log-Odds	Odds Ratio	Log-Odds	Odds Ratio	Log-Odds	Odds Ratio	Log-Odds	Odds Ratio
Formation Coefficients								
Tie Likelihood	0.78 (1.04)	2.19	-3.48 (0.87) *	0.03	-0.84 (0.82)	0.43	-0.82 (0.71)	0.44
Elapsed Time in Isolation	-0.26 (0.05) *	0.77	-0.10 (0.02) *	0.91	-0.18 (0.03) *	0.83	-0.12 (0.02) *	0.89
Time Between Observations	-0.02 (0.09)	0.98	0.22 (0.08) *	1.24	0.13 (0.07) ˙	1.14	0.13 (0.06) *	1.14
Reciprocity	1.12 (0.78)	3.06	0.24 (0.50)	1.28	0.65 (0.56)	1.91	0.40 (0.44)	1.50
Race Homophily	1.49 (0.60) *	4.43	-0.05 (0.35)	0.95	0.79 (0.42) ˙	2.19	-0.16 (0.33)	0.85
Military Experience Homophily	0.24 (0.53)	1.27	0.55 (0.36)	1.74	0.55 (0.40)	1.74	0.14 (0.33)	1.15
Dissolution Coefficients								
Tie Likelihood	3.35 (1.12) *	28.64	-1.82 (1.10) ˙	0.16	2.98 (1.11) *	19.70	2.12 (0.78) *	8.35
Elapsed Time in Isolation	0.04 (0.07)	1.04	-0.03 (0.09)	0.97	0.09 (0.06)	1.09	0.02 (0.05)	1.02
Time Between Observations	-0.06 (0.09)	0.94	0.08 (0.12)	1.08	-0.19 (0.09) *	0.82	-0.08 (0.08)	0.93
Reciprocity	-0.01 (0.71)	0.99	-0.73 (1.10)	0.48	0.12 (0.60)	1.13	0.43 (0.49)	1.54
Race Homophily	-0.82 (0.58)	0.44	1.69 (0.86) *	5.39	-0.14 (0.50)	0.87	-0.60 (0.45)	0.55
Military Experience Homophily	-0.90 (0.56)	0.41	2.25 (0.93) *	9.48	-0.32 (0.53)	0.73	0.06 (0.44)	1.06
AIC	178.9		89.64		182.8		193.7	
BIC	203.3		105.1		206.8		216.3	

Standard error in parentheses. * p<0.05, ˙ p <0.10

Table 3. Correlation testing results for each social relation

TASK AFFECT NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	0.30 (0.37)	-0.30 (0.08) ·	0.15 (0.00) *	0.45 (0.00) *
Normalized Transitivity	-0.27 (0.29)	0.22 (0.33)	-0.17 (0.05) *	-0.38 (0.08) ·
Normalized Cyclicality	-0.20 (0.39)	0.27 (0.30)	0.31 (0.23)	0.47 (0.00) *
Closeness Centralization	-0.33 (0.00) *	0.23 (0.00) *	-0.13 (0.24)	-0.43 (0.00) *
Relative Indegree of Commander	0.28 (0.12)	0.03 (0.00) *	-0.19 (0.41)	-0.04 (0.01) *
Relative Outdegree of Commander	-0.02 (0.02) *	-0.27 (0.50)	0.42 (0.00) *	0.26 (0.08) ·
Normalized Two-Paths	-0.25 (0.50)	0.33 (0.06) ·	0.17 (0.77)	-0.23 (0.02) *

TASK HINDRANCE NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	-0.23 (0.18)	0.30 (0.01) *	0.13 (0.24)	-0.34 (0.06) ·
Normalized Transitivity	-0.13 (0.34)	0.11 (0.21)	0.00 (0.51)	-0.17 (0.15)
Normalized Cyclicality	-0.28 (0.10)	0.10 (0.20)	0.27 (0.17)	0.02 (0.35)
Closeness Centralization	-0.19 (0.18)	0.08 (0.16)	-0.14 (0.35)	-0.25 (0.03) *
Relative Indegree of Commander	0.01 (0.72)	0.00 (0.92)	0.38 (0.01) *	0.26 (0.09) ·
Relative Outdegree of Commander	0.03 (0.70)	0.36 (0.00) *	-0.30 (0.21)	-0.54 (0.00) *
Normalized Two-Paths	0.25 (0.05) ·	0.30 (0.03) *	0.07 (0.31)	-0.19 (0.19)

LEADERSHIP NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	0.35 (0.17)	-0.40 (0.01) *	0.21 (0.00) *	0.35 (0.00) *
Normalized Transitivity	-0.41 (0.04) *	0.28 (0.26)	-0.20 (0.03) *	-0.20 (0.27)
Normalized Cyclicality	0.20 (0.03) *	-0.32 (0.00) *	-0.08 (0.22)	0.44 (0.00) *
Closeness Centralization	-0.43 (0.00) *	0.42 (0.00) *	-0.15 (0.41)	-0.33 (0.00) *
Relative Indegree of Commander	0.36 (0.05) *	-0.33 (0.28)	-0.10 (0.71)	0.20 (0.10)
Relative Outdegree of Commander	-0.05 (0.00) *	0.40 (0.00) *	-0.14 (0.65)	-0.31 (0.00) *
Normalized Two-Paths	0.08 (0.00) *	-0.39 (0.00) *	0.09 (0.82)	0.50 (0.00) *

FOLLOWERSHIP NETWORKS				
	Performance Measures			
	Generate	Choose	Negotiate	Execute
Network Properties				
Density	0.03 (0.76)	-0.40 (0.04) *	0.36 (0.00) *	0.38 (0.00) *
Normalized Transitivity	-0.05 (0.56)	0.14 (0.36)	0.00 (0.69)	-0.13 (0.36)
Normalized Cyclicality	-0.19 (0.22)	-0.37 (0.03) *	0.36 (0.03) *	0.32 (0.05) ·
Closeness Centralization	-0.03 (0.04) *	0.18 (0.00) *	-0.03 (0.88)	-0.22 (0.00) *
Relative Indegree of Commander	-0.21 (0.00) *	0.36 (0.00) *	-0.44 (0.02) *	-0.32 (0.00) *
Relative Outdegree of Commander	0.05 (0.89)	-0.08 (0.97)	0.20 (0.00) *	0.18 (0.19)
Normalized Two-Paths	0.04 (0.14)	-0.29 (0.01) *	0.28 (0.25)	0.16 (0.09) ·

P-value in parentheses. * $p<0.05$, · $p<0.10$

Correlation Significance Testing. Table 2 presents the parameter estimates from separate STERGM models for each network, which were used to perform sampling from the null distribution. Table 3 reports the Spearman rank correlation coefficient between each of the network statistics and performance, alongside p-values for each, calculated using our method to generate a null distribution from 250 simulated correlation coefficients. For brevity, we describe the network relation with the strongest association with each performance dimension.

Naturally, multiple testing problems [1] occur when performing a large quantity of significance tests. Since we intended this as an exploratory analysis, we did not account for multiple testing effects here. For stricter hypothesis testing, p-values should be adjusted using an approach such as Bonferroni correction [1].

For task affect ties, closeness centralization is inversely related to performance on the Generate task, and positively related to performance on the Choose task. Relative outdegree of the commander is positively associated with performance on the Negotiate task, while cyclicality is positively related to performance on the Execute task. For task hindrance ties, no network statistics were significantly related to performance on the Generate task. However, hindrance density was positively related to Choose and Execute task performance. Finally, the commander's relative indegree is positively related to Negotiate task performance.

For leadership ties, closeness centralization is inversely related to performance on the Generate task, but positively related to performance on the Choose task. Leadership density is positively related to performance on the Negotiate task, whereas leadership two-paths are positively associated with performance on the Execute task. For followership ties, the relative indegree of the commander is inversely related to performance on the Generate, Negotiate, and Execute tasks, while followership density is positively related to Choose task performance.

5 Discussion

These findings illustrate how STERGM can be used to account for endogeneity due to time when correlating network statistics with an exogenous network level outcome variable. This approach considers the association between network properties and team performance by decomposing the network into individual ties to be modeled. The STERGM-based sampling from the null distribution controls for complex interdependencies between ties (e.g. reciprocity, closure, or tendency of ties to persist over time). We model the network statistic not as a single continuous variable, but as a consequence of a number of discrete ties that have complex interdependencies with one another. STERGM therefore provides a flexible framework to control for various types of interdependencies that have been well-established as occurring across many real-world social networks.

We apply simulation from the null distribution to answer the question: Which network patterns predict team performance? The results suggest elements of closure, centralization, and subgrouping along different relations affect performance. Additionally, we observed multiple cases where a structure that benefited one type of performance undermined another. Further work is needed to discover the underlying social dynamics linking networks to team performance.

Correlations on small sample sizes, such as ours, are difficult to interpret because of the potential for spurious findings. The approach we employ for generating an empirical p-value, which takes into account temporal and other structural dynamics, provides a means for understanding the probability of finding such an effect. In doing so, it serves as a tool to help interpret effects when working with a moderate to small sample of networks.

Though we demonstrate a tool for statistical testing on small network samples, it has a number of limitations. First, the method needs to be compared to existing multilevel techniques which also account for temporal endogeneity [2,11,16]. Second, this approach needs to be explored as it applies to smaller or larger samples. How few measurements would merit this approach, and how many measurements could it be usefully applied to? Simulation studies could help explore these questions.

Acknowledgements. This material is based upon work supported by NASA under award numbers NNX15AM32G, NNX15AM26G, and 80NSSC18K0221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

References

1. Aickin, M., Gensler, H.: Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. Am. J. Public Health **86**(5), 726–728 (1996)
2. Bell, S.T., Fisher, D.M., Brown, S.G., Mann, K.E.: An approach for conducting actionable research with extreme teams. J. Manage. **44**(7), 2740–2765 (2018)
3. Crawford, E.R., LePine, J.A.: A configural theory of team processes: accounting for the structure of taskwork and teamwork. AMRO **38**(1), 32–48 (2013)
4. Freeman, L.C.: Centrality in social networks conceptual clarification. Soc. Netw. **1**(3), 215–239 (1978)
5. Krackardt, D.: QAP partialling as a test of spuriousness. Soc. Netw. **9**(2), 171–186 (1987)
6. Krackhardt, D.: Predicting with networks: nonparametric multiple regression analysis of dyadic data. Soc. Netw. **10**(4), 359–381 (1988)
7. Krivitsky, P.N., Handcock, M.S.: A separable model for dynamic networks. J. R. Stat. **76**(1), 29–46 (2014)
8. Krivitsky, P.N., Handcock, M.: tergm: Fit, Simulate and Diagnose Models for Network Evolution Based on Exponential-Family Random Graph Models. The Statnet Project (https://statnet.org) R package version 3 (0) (2019)
9. Landon, L.B., Slack, K.J., Barrett, J.D.: Teamwork and collaboration in long-duration space missions: going to extremes. Am. Psychol. **73**(4), 563 (2018)
10. Larson, L., Wojcik, H., Gokhman, I., DeChurch, L., Bell, S., Contractor, N.: Team performance in space crews: Houston, we have a teamwork problem. Acta Astronautica **161**, 108–114 (2019)
11. Lazega, E., Snijders, T.A.: Multilevel Network Analysis for the Social Sciences: Theory, Methods and Applications, vol. 12. Springer, Cham (2015)

12. Lungeanu, A., Carter, D.R., DeChurch, L.A., Contractor, N.S.: How team interlock ecosystems shape the assembly of scientific teams: a hypergraph approach. Commun. Methods Meas. **12**(2–3), 174–198 (2018)
13. McGrath, J.E.: Groups: Interaction and Performance, vol. 14. Prentice-Hall, Englewood Cliffs (1984)
14. Mukherjee, S., Uzzi, B., Jones, B., Stringer, M.: A new method for identifying recombinations of existing knowledge associated with high-impact innovation. J. Prod. Innov. Manag. **33**(2), 224–236 (2016)
15. Salas, E., Tannenbaum, S.I., Kozlowski, S.W., Miller, C.A., Mathieu, J.E., Vessey, W.B.: Teams in space exploration: a new frontier for the science of team effectiveness. Curr. Dir. Psychol. Sci. **24**(3), 200–207 (2015)
16. Snijders, T.A.: Multilevel Analysis. Springer, Heidelberg (2011)
17. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. **15**(1), 72–101 (1904)