Visual Inference for a Social Network Model

Samantha Tyner*
Department of Statistics and Statistical Laboratory, Iowa State University
and
Heike Hofmann
Department of Statistics and Statistical Laboratory, Iowa State University

July 23, 2019

Abstract

Three of the most important assessments of a statistical model are significance tests of parameters, goodness-of-fit tests, and power calculations of the tests. All three tasks become more difficult as the model becomes more complex. We will explore these three assessments of one particularly complex set of models, continuous time Markov chain (CTMC) models, for dynamic social networks (Snijders, 1996). In this paper, we propose new methods for significance and goodness-of-fit testing, as well as power calculations for CTMC models via the visual inference (VI) paradigm of Buja et al. (2009). With VI, we can look at entire datasets simulated from a model, instead of relying a single metric such as a p-value. We conducted a VI experiment, with participants recruited via Amazon Mechanical Turk, to assess the significance of CTMC parameters, the fit of a CTMC model, and the visual power of CTMC parameters, and found that ... TBD

Keywords: social network analysis, visual inference, dynamic networks, network visualization, network mapping, goodness-of-fit, hypothesis testing

*The authors gratefully acknowledge funding from the National Science Foundation Grant # DMS 1007697. All data collection has been conducted with approval from the Institutional Review Board IRB 10-347

1 Background

When selecting and fitting statistical models, there are typically three key assessments of interest: significance tests of parameters, goodness-of-fit tests, and power calculations. With significance testing, the null hypothesis assumes that the data come from a simple model nested within the model of interest. We then conduct significance tests of one or more additional parameters to determine how much of the variability in the data they explain. For goodness-of-fit tests, we examine one or more models of interest to assess how well these models explain the data. Power then quantifies the ability of the hypothesis test to detect the difference between the null and alternative hypothesis. All three of these elements of statistical modeling are vital to ensure that we can draw valid conclusions from a model.

The more complex the model, however, the harder it typically is to assess significance, power, and goodness-of-fit. One particularly complicated family of models are those designed to model networks. A *network* is any set of things, such as people, or computers, that are connected in some way, through social relations or the Internet. In a social network, the people are the *nodes* or *actors*, and the relationships between the people are *edges* or *ties*. In this paper we use nodes and edges interchangeably with actors and ties, respectively. Models for networks are particularly complex, as dependencies inherent to network data make them difficult to model. This difficulty increases further when studying *dynamic networks*, the same set of nodes and their changing relationships observed at many points in time, because of the added temporal dimension. Yet, researchers are often interested in modeling dynamic social networks, such as friendship networks among students or collaboration networks between legislators. Even some of the simplest network models, however, lack the asymptotics required to perform traditional goodness-of-fit tests (Holland and Leinhardt, 1981). In addition, a direct maximum likelihood estimation of model parameters is frequently impossible due to the intractability of the models (Hummel et al., 2012).

In order to circumvent some of these difficulties, we propose a new approach for significance testing of parameters, goodness-of-fit testing, and power calculations for one family social network models: continuous time Markov-chain (CTMC) models for dynamic network data, as described in Snijders (1996). We use *visual inference* in place of traditional statistical methods for social network models, such as t-tests for significance of parameters or outdegree distributions for goodness-of-fit tests. Visual inference (VI), introduced by Buja et al. (2009), allows us to look at the *entire* dataset simulated from a network model, whereas traditional methods use one-dimensional metrics derived from the network or a p-value for a parameter in the model. By using VI to supplement traditional statistical tests, we gain insight into the role of the parameters in these CTMC models, and we gain the ability to assess the fit of the CTMCs to dynamic network data.

The paper is outlined as follows: Section 1.1 provides an introduction to the CTMC family of models. Section 1.2 gives a basic overview of visual inference and the lineup protocol. In Section 2, we describe our example data and define our models of interest that we used to develop our VI methods. Section 3 details how we designed a VI experiment for significance testing and goodness of fit procedures for CTMC models, and Section 4 details the results of our experiment. We close with a discussion in Section 5.

1.1 CTMC Models for Dynamic Social Network Data

We define CTMC models for dynamic social network data here in their barest form. Full details on these models can be found in Snijders (1996, 2001); Snijders et al. (2010a,b, 2007); Snijders (2017). CTMCs are a family of models for dynamic network data that incorporate both network structure and node-level covariates to describe how a network changes over time (Snijders, 1996). Traditional network models, such as exponential random graph models, usually only consider network structure. Social networks are ever-changing as relationships decay or grow over time, and each actor in a network has characteristics that affect ties to other actors in the network. CTMC models use the network structure and the node covariate information, and as such can be very complicated models. As network data are inherently complex, it can be difficult to interpret model parameters and their estimates. There are also many possible parameters to include in a CTMC, which makes parameter selection and goodness-of-fit testing challenging.

CTMC models use network structure and node covariates to model the network change one tie at a time. There are two model pieces: a *rate function* that dictates *how often* changes in the network occur and an *objective function* that determines *what* those changes are.

Rate Function: All changes in the network are treated as choices made by the actors in the network. The rate function dictates when changes are made and which actor can make them. When chosen, the actor, i, gets a chance to make a change to one of the other nodes j. In general, the rate function can include structural and node covariate parameters into account so that each actor has a different rate of change. However, we choose a simple rate function that is constant over all nodes in a given time period, because we focus on interpreting the parameters of the objective function. We denote the rate from t_m to t_{m+1} as α_m for $m = 1, \dots, M-1$. Using this notation, the *waiting time* to the next chance for actor i to make a change is exponentially distributed with expected value α_m^{-1}. Since the rate is the same for all actors, the waiting time for *any* actor to get the opportunity to change its set of ties is also exponentially distributed with expected value $(n\alpha_m)^{-1}$.

Objective Function: After actor i has been selected by the rate function to change, it randomly picks one of its current ties, x_{ij}, to change. The probability that actor i changes its current tie to actor j is determined by the *objective function* of the model and a random error term U with log-Weibull distribution (Snijders, 2005). Actor i aims to maximize the objective function f_i given the current state of the network, x and the node-level covariates, $\mathbf{Z}$. This function is defined as:

$$f_i(x, \boldsymbol{\beta}, \mathbf{Z}) = \sum_{k=1}^{K} \beta_k s_{ik}(x, \mathbf{Z}), \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)$ are additional model parameters, each associated with some statistics, $s_{i1}(x, \mathbf{Z}), \dots, s_{iK}(x, \mathbf{Z})$, calculated for actor i at the current network state x. TAt least two parameters must be included in the objective function: density and reciprocity (Ripley et al., 2017). We denote the density, or out-degree, parameter by β_1 and the associated statistic as $s_{i1}(x)$ and we denote the reciprocity parameter by β_2 and the associated statistic as $s_{i2}(x) = \sum_j x_{ij}x_{ji}$. We will refer to the very simple model with only these two parameters in the objective function as model M1. We define additional parameters and models of interest in Section 2 Version 1.2-3 of `RSiena` (Ripley et al., 2013), the software we use to fit CTMC models to data, provides over 80 possible effects that can be included in the objective function.

The objective function $f_i(x, \boldsymbol{\beta}, \mathbf{Z})$ and random component U combine to form the *transition probability*, p_{ij} of the network changing from its current state x to the state $x(i \rightsquigarrow j)$, which is identical to x except for x_{ij}: $x_{ij}(i \rightsquigarrow j) = 1 - x_{ij}$. The transition probability is

$$p_{ij} = \frac{\exp\{f_i(x(i \rightsquigarrow j), \boldsymbol{\beta}, \mathbf{Z})\}}{\sum_h \exp\{f_i(x(i \rightsquigarrow h), \boldsymbol{\beta}, \mathbf{Z})\}}, \tag{2}$$

dictating which edge node i changes.

1.2 Visual Inference

Data visualizations are an important component of data analysis, providing a mechanism for discovering patterns in data. Pioneering research by Gelman (2004), Buja et al. (2009) and Majumder et al. (2013) provide methods to quantify the significance of discoveries made from visualizations. Buja et al. (2009) introduced two protocols, the Rorschach and the lineup protocol, which bridge the gulf between traditional statistical inference and exploratory data analysis. Here, we use the lineup protocol to design significance, goodness-of-fit (GoF) and power tests. Under the protocol, we begin with a data set of interest to us, such as a network, a visualization of this data, such as a node-link diagram, and a model of interest. We consider two hypotheses: the null hypothesis that the model of interest generated the data, and the alternative hypothesis that the data were not generated under this model. To construct a lineup of size P, $P - 1$ sets of data are simulated from the null model. Each of the $P - 1$ simulated datasets are visualized in the same way as the data, and the plot of the data is placed randomly among the set of $P - 1$ *null plots*. Human observers then examine the lineup and identify the plot(s) that look(s) most different from the others. If an observer identifies the data plot, this is evidence against the null hypothesis. An observer has a chance of P^{-1} to pick the data plot from the lineup by simply guessing. The evidence grows in strength with the number of independent observers identifying the data plot.

The lineup protocol places a *plot* in the framework of hypothesis tests: the plot of the data is the test statistic, which is compared against the null plots, representing the sampling distribution under the null hypothesis. The lineup protocol was formally tested for linear models in a head-to-head comparison with the equivalent conventional test in Majumder et al. (2013). The experiment utilized human subjects from Amazon's Mechanical Turk (Amazon, 2010) and used simulation to control conditions. The results suggest that VI done in a controlled setting gives similar results to conventional tests. This is evidence that VI can be used when no conventional tests exist or when testing is difficult.

2 Example Data and Models

The data we use are collaboration networks in the United States Senate during the 111^{th} through 114^{th} Congresses. These senates began on January 6, 2009, the start date of the 111^{th} Congress, and ended on January 3, 2017, the last date of the 114^{th} Congress Details of how this data can be downloaded are provided by François Briatte at `github.com/briatte/congress`. In the US Senate, senators often show support for a piece of legislation by co-sponsoring a bill authored by one of their colleagues. In a co-sponsorship network, ties are directed from senator i to senator j when senator i signs on as a co-sponsor to the bill that senator j authored. There are many hundreds of ties between senators when they are connected in this way, so we simplify the network by computing a single value for each senator-senator collaboration called the *weighted propensity to co-sponsor* (WPC). This value is defined in Gross et al. (2008) as

$$WPC_{ij} = \sum_{b=1}^{B_j} \frac{Y_{ij(b)}}{c_{j(b)}} \left(\sum_{b=1}^{B_j} \frac{1}{c_{j(b)}} \right)^{-1} \tag{3}$$

where B_j is the number of bills in a congressional session authored by senator j, $c_{j(b)}$ is the number of co-sponsors on senator j's b^{th} bill, where $b \in \{1, \ldots, B_j\}$, and $Y_{ij(b)}$ is an indicator variable that senator i co-sponsored senator j's b^{th} bill. This measure ranges in value from 0 to 1, where $WPC_{ij} = 1$ if senator i is a co-sponsor on every one of senator j's bills and $WPC_{ij} = 0$ if senator i is never a co-sponsor any of senator j's bills. Because CTMC models require binary edges, we construct the edges as follows:

$$x_{ij} = \begin{cases} 1 & WPC_{ij} > 0.25 \\ 0 & WPC_{ij} \leq 0.25 \end{cases} \tag{4}$$

so that only strong relationships between senators are examined. For each of the four senate sessions, we have three node covariates: the party affiliation of each senator, the number of bills they authored in each session, and their gender. We use each of these covariates in CTMC models to try to explain how ties are formed between senators over time. The node-link diagram for the senate network is shown in Figure 1. We labelled some of the nodes in the network whose names we think will be familiar the reader, either because they are leaders in their party or they have run for president. The size of the nodes represents how many bills the senator authored in a session, the color represents party affiliation, and the shape represents gender. In each of the four sessions, there is one very large connected component tying many of the prominent senators together, with many smaller connected groups surrounding the larger component. In each senate, the structure changes slightly as new senators arrive or come to prominence.

In legislative co-sponsorship networks, it is known that party affiliation, reciprocity of relationships, and whether senators are "workhorses" who author many bills or "show horses" who author few bills, are major influences on structure (Ringe et al., 2016). We focus on these covariates, plus the additional gender covariate when choosing which CTMC models to fit to the data.

Models: We examine a total of six models, each identified by its objective function. The parameters in the objective function, $\beta_1, \ldots, \beta_6$ are defined in Table 1. All models have the outdegree and reciprocity parameters, β_1 and β_2, included. Other effects were added one at a time so that the simplest model $M1$ is nested in each of the other models. The largest model is M7: each of the other five models is nested in it. We used Wald tests in `RSiena` as described in Section 4 to perform significance tests on $\beta_3, \ldots, \beta_6$ in models M3, ..., M6. The most significant effect is β_3, the jumping transitive triplet (JTT) parameter for the party covariate, which was estimated to be about -5.9 with a standard error of 0.11, resulting in a Wald p-value of less than 0.0001. This parameter considers the number of transitive closures formed between two senators from different parties. The large negative estimate is an indication that forming transitive ties between two people from different parties is strongly discouraged, which comports with the divisive nature of American politics. Another significant effect is β_4, the same JTT parameter for the sex covariate, with an

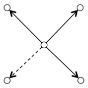
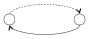
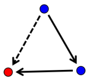
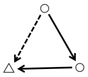
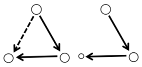
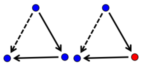
β_k	Effect name	Interaction Variable	Formula	Picture	Initial estimate	Wald p-value
β_1	density	–	$s_{i1}(x) = \sum_j x_{ij}$		2.204	NA
β_2	reciprocity	–	$s_{i2}(x) = \sum_j x_{ij}x_{ji}$		-4.903	NA
β_3	jumping transitive triplet	party	$s_{i3}(x, \mathbf{p}) = \sum_{j \neq h} x_{ij}x_{ih}x_{hj} \cdot \mathbb{I}(p_i = p_h \neq p_j)$		-5.884	< 0.0001
β_4	jumping transitive triplet	sex	$s_{i4}(x, \mathbf{s}) = \sum_{j \neq h} x_{ij}x_{ih}x_{hj} \cdot \mathbb{I}(s_i = s_h \neq s_j)$		3.335	0.0002
β_5	similarity transitive triplet	bills	$s_{i5}(x, \mathbf{b}) = \sum_j x_{ij}x_{ih}x_{hj} \cdot (sim^b_{ij} - \overline{sim}^b)^*$		9.821	0.0128
β_6	same transitive triplet	party	$s_{i6}(x, \mathbf{p}) = \sum_j x_{ij}x_{ih}x_{hj} \cdot \mathbb{I}(p_i = p_j)$		1.306	0.0642

Table 1: The effects we used in the CTMC models fit to the senate data. In the picture which represents each effect, the dotted tie is encouraged to form if the estimate is positive, and discouraged to form if the estimate is negative. * - sim^b_{ij} is defined in Equation 5 and $\overline{sim}^b = \frac{1}{n(n-1)} \sum_{i \neq j} sim^b_{ij}$ is the average bill similarity score between two senators.

Model	β_1	β_2	β_3	β_4	β_5	β_6
M1	✓	✓	–	–	–	–
M3	✓	✓	✓	–	–	–
M4	✓	✓	–	✓	–	–
M5	✓	✓	–	–	✓	–
M6	✓	✓	–	–	–	✓
M7	✓	✓	✓	✓	✓	✓

Table 2: The models we use defined by the parameters in their objective functions. The corresponding parameter for s_{ik} is β_k. Note there is no model M2.

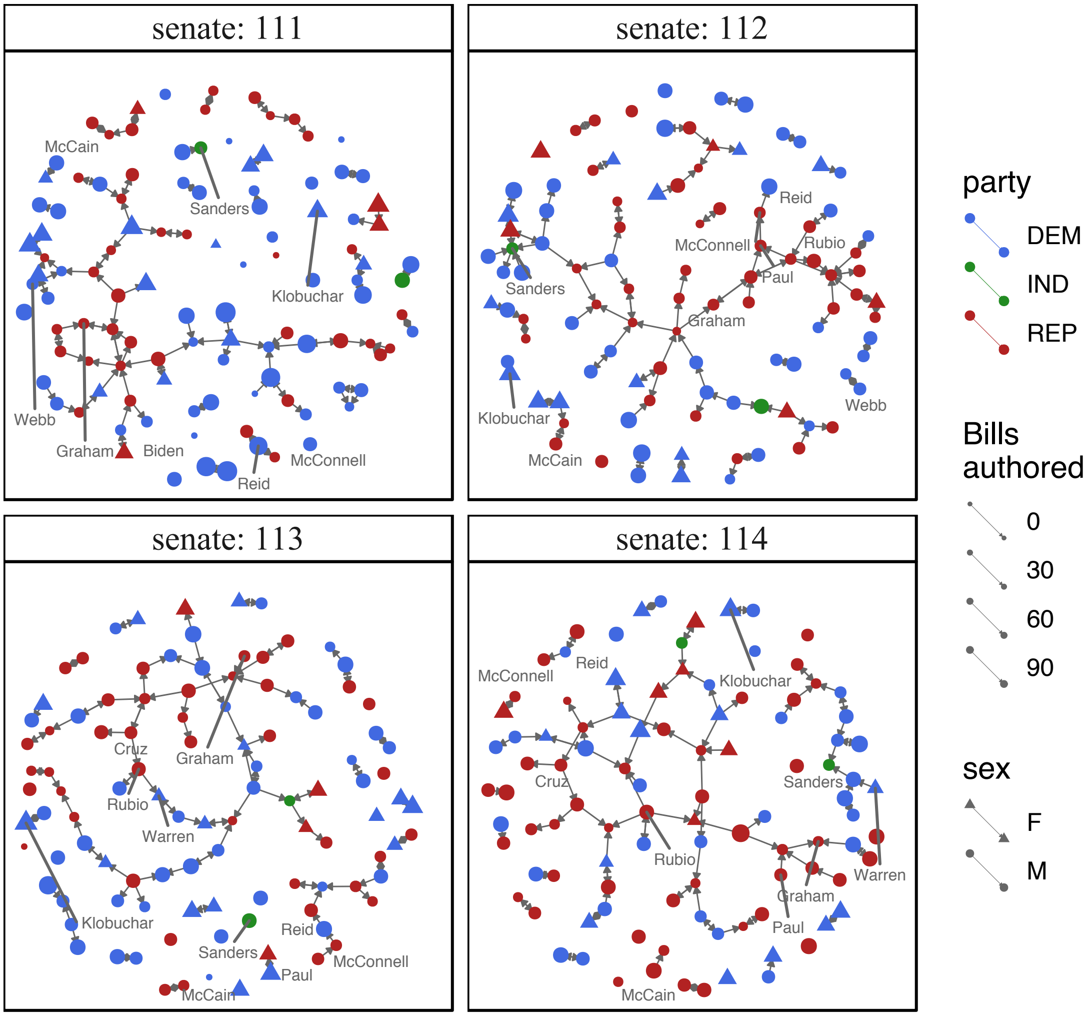


Figure 1: The US senate collaboration network observed at four time points. Color represents party, shape represents gender, and size represents number of bills authored in a session. The Fruchterman-Reingold layout is shown (Fruchterman and Reingold, 1991). Drawn with the `geomnet` R package. (Tyner and Hofmann, 2018)

estimate of about 3.3 with a standard error of 0.89. This parameter also considers transitive closures, but for senators of different genders. The positive value indicates that transitive ties between senators of different genders are more likely to form. Next, we consider β_5, the covariate-related similarity score-weighted transitive triplets parameter for the number of bills authored by a senator. We chose to look at similarity instead of raw covariate value because the number of bills authored is more continuous than gender or party. The similarity measure is computed as:

$$sim^b_{ij} = \frac{\max_{hk} |b_h - b_k| - |b_i - b_j|}{\max_{hk} |b_h - b_k|} \tag{5}$$

where $\max_{hk} |b_h - b_k|$ is the range of number of bills authored by senators, and b_i, b_j are the number of bills authored by senators i, j respectively in the senate period. This effect was estimated at about 9.8 with standard error of 3.9. The high positive estimate suggests senators are encouraged to collaborate with other senators who author about the same number of bills they do. This tendency of senators to co-sponsor bills written by senators who are similarly "prolific" corresponds to the tendency of senators to be either "workhorses" or "show horses". Senators known as workhorses author many pieces of legislation in a session, and largely stay out of the public arena. The show horse senators, on the other hand, author relatively few pieces of legislation, and tend to appear in the media very frequently. Finally, we found β_6, the same party transitive triplet effect to be significant, with a fitted value of 1.3 and standard error of 0.7, meaning that transitive relationships between senators tend to form when they are from the same party, exactly as we would expect in a legislative body in a country with deeply entrenched partisan divides.

We fit all six of our models in `RSiena` using Markov Chain Monte Carlo (MCMC) methods to approximate the method of moments estimates of the parameters. Because the estimation is done through MCMC simulation, we fit each model to the data 1,000 times. From the simulations in which the MC converged, which made up over 90% of the fits for each model, we computed the mean of the estimates of each parameter to get final estimates of $\hat{\boldsymbol{\beta}}$ for each model, which are shown in Table 3.

We want to determine the role that each of these parameters plays in the objective functions for the different models. So, we use the estimates given in Table 3 to simulate from models M1 through M6. We discuss the simulation procedure and how we use the simulations in Section 3.

3 Experiment Design

We want to assess CTMC models for social network data using the lineup protocol in three ways: (1) significance tests of parameters, (2) goodness-of-fit tests of a model, and (3)

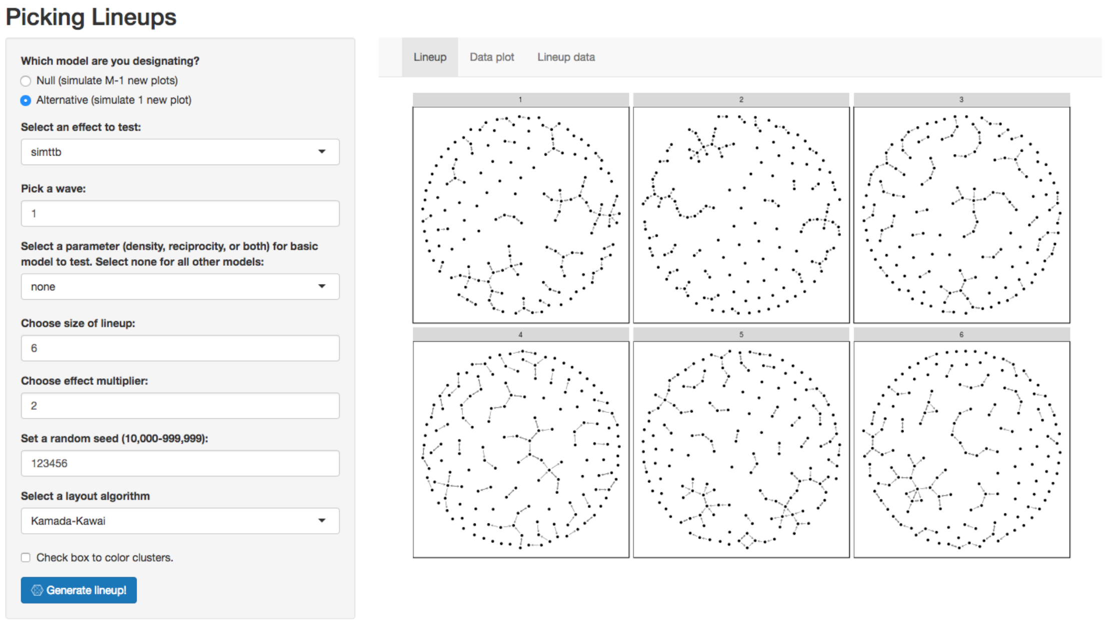
Model	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
M1	2.441	2.46	2.204	-4.903	4.893	–	–	–	–
M3	2.44	2.46	2.204	-4.902	4.893	-3.45	–	–	–
M4	2.438	2.461	2.211	-4.918	4.898	–	3.34	–	–
M5	2.442	2.459	2.206	-4.917	4.89	–	–	10.091	–
M6	2.443	2.461	2.205	-4.911	4.881	–	–	–	1.329
M7	2.441	2.459	2.21	-4.923	4.892	–	2.374	6.966	0.205

Table 3: The final estimates from repeated estimation of our models of interest. When simulating from these models, these are the estimates that we will use unless otherwise stated.

determining visual power of the effects. Each one of these situations requires a different setup, which we describe in detail, making use of the lineup protocol defined in Buja et al. (2009). In each lineup, we include plots from two models: a null model and an alternative model. The definition of the null and alternative model varies with the model and the assessment we explore.

Typically, a lineup shows sets of 20 plots at a time, for example in Loy et al. (2015); Vander Plas and Hofmann (2015), but we determined that looking at 20 node-link diagrams at once is too difficult. We chose to present our participants with only six plots at a time in order to show the node-link diagrams in more detail and to reduce cognitive load. To construct a lineup, we simulate five networks from the null model and one network from the alternative model. Several lineups shown to our participants are presented and discussed in Section 4.

To simulate lineups from the models we used the `siena07` function in `RSiena` (Ripley et al., 2013). For the purposes of our experiment, we focus on simulating the second "wave" of data, the 112^{th} Senate network, and we condition on the first wave of data, 111^{th} Senate network. Sections 3.1 through 3.3 describe in detail how we constructed the lineups, which parameter values we simulate from, and why. Lineups we created were shown to independent observers recruited through Amazon Mechanical Turk for feedback, and we provide more detail on the Turk setup in Section 4.



Picking Lineups

Which model are you designating?

○ Null (simulate M-1 new plots)

● Alternative (simulate 1 new plot)

Select an effect to test: simttb

Pick a wave: 1

Select a parameter (density, reciprocity, or both) for basic model to test. Select none for all other models: none

Choose size of lineup: 6

Choose effect multiplier: 2

Set a random seed (10,000-999,999): 123456

Select a layout algorithm: Kamada-Kawai

□ Check box to color clusters.

Figure 2: A screen shot of the web application we created to help design our lineup experiment. More details about this application are given in Section 3.3. In the lineup, M5 is the alternative model with β_5 set to twice its estimated value given in Table 3. One plot simulated from this model is placed at random among five observations simulated from the null model, M1.

3.1 Significance Testing

In the significance testing protocol, a parameter of interest β_k is selected to test. The hypotheses we use to generate lineups are:

$$H_0 : \beta_k = 0 \quad \text{versus} \quad H_A : \beta_k \neq 0 \tag{6}$$

Under the null hypothesis, we assume that the model that generated the network data is M1, the simplest model presented in Section 2. In the lineup, there are five null plots, constructed from five simulations from M1 of the second wave, with β_1, β_2 set to the corresponding estimates given in Table 3. The alternative model is the model with β_1, β_2, and β_k in the objective function. The alternative data plot is simulated from the appropriate model with values set to the estimates from Table 3. If an observer picks out the alternative data plot, that is evidence against the null hypothesis, while picking one of the null plots is evidence in favor of the null hypothesis. To avoid over-working our participants, we chose to test only two parameters, β_3 and β_4. These two had the smallest p-values from the significance tests, so we chose them for the visual significance test as well because we hypothesized they would be easier to pick out of the lineup. Thus, we compare the null model M1 to the alternative models M3 and M4, using three repetitions of each hypothesis test in the experiment.

3.2 Goodness-of-Fit Testing

For the goodness-of-fit tests, we compare one model of interest to the data. The hypothesis we use to generate lineups are

H_0: The data come from the model of interest

H_A: The data come from some other, unknown model

To generate the null plots, we simulate five networks the model of interest using the corresponding parameters in Table 3. We pick a wave to focus on, wave two, which is the first simulated network, and among these five plots, we place a node-link diagram of the true second wave of data. We cannot show the data more than once to each participant. The models we chose for goodness-of-fit testing are M3, M4, M5, and M7, and each participant only saw one lineup for the goodness-of-fit tests.

3.3 Visual Power Testing

Using VI, we want to determine at what value an additional effect included in the model becomes noticeable. By *noticeable*, we mean that the inclusion of the effect alters the appearance of networks simulated from the model so much that many independent viewers are able to pick out the one node-link diagram rendered from data simulated from the model *with* the effect in a lineup among five plots simulated from the model *without* the effect. In this way, we measure the visual power of a parameter. We perform visual power tests for all parameters in the objective function, $\beta_1, \dots, \beta_6$.

In model M1, with only two parameters in the objective function, we varied both the density and reciprocity parameter values one at a time, keeping all other parameters at their fitted values given in Table 3. In models M3 through M6, we vary the additional parameter, β_3 through β_6, respectively, while holding all other parameters at their value in Table 3. We want to determine how the size of these parameters affects the overall structure of the network data simulated from the models M1 through M6, so we also vary the magnitude of the parameters in order to determine at what value the effects become noticeable.

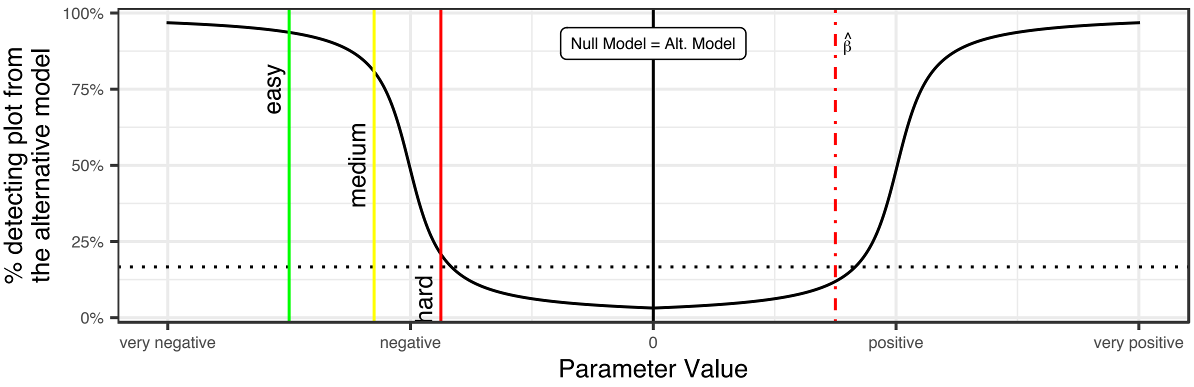


Figure 3: As the parameter increases in absolute value, more viewers of the lineup should pick the alternative data out of a lineup. Note that the significance test we construct in Section 3.1 is just one point on the curve, represented by the vertical dotted line labeled $\hat{\beta}$. The easy, medium, and hard lines represent how we determined which values of the parameters to show to our participants, and the horizontal dotted line shows the type-I error for one viewer of a lineup of size 6.

To determine when an effect becomes noticeable, we examine six different levels of the effect, three negative and three positive ones. Figure 3 shows a sketch of what the selection probability looks like hypothetically with varying effect size: the higher in absolute value the parameter is, the more likely participants are to select the alternative model out of the lineup. To determine the exact values of the six levels we want to test for each effect, we started with the estimates of the parameter at hand (see Table **??**), and used small negative and positive factors to determine at what point *we* noticed the effect of the parameter in simulations from the changed models. In Figure 3, we demonstrate these values with the vertical lines labeled "easy," "medium," and "hard." We expect most viewers to see the effects at the "easy" values, and we expect very few, if any, viewers to see the effects at the "hard" values.

To decide on the values to use for each difficulty level, we constructed an online application that simulated the lineup protocol for us to be the guinea pigs in our own experiment (Swan, 2013). A screen shot of the app we created with the `shiny` package by Chang et al. (2017) is shown in Figure 2. On the left side of the screen, the user[1] can input the information necessary for creating a lineup from the models M1 through M7 for the data in Section 2:

1. Choose to simulate one plot to change the alternative model in the lineup, or simulate $P - 1$ plots to change the null model in the lineup
2. Choose the model from which to simulate and choose the wave of data to simulate.
3. If M1 is selected in (2), select whether to alter the density (β_1) or the reciprocity (β_2) parameter.
4. Choose the size of the lineup, P, the magnitude of the effect size, a random seed for reproducibility, and a layout algorithm to use for the node-link diagrams. There is also a checkbox if the user wishes the nodes to be colored by the size of the connected component to which they belong.

All plots that are *not* from the model specified by steps 1-4 above are simulated from model M1 with the estimates of the rate parameters, and β_1 and β_2 as given in Table 3.

Using our Shiny application, we settled on six parameter values to test for each of our six effects, $\beta_1, \dots \beta_6$. All values of the parameters used in the experiment are given in Table 4. In the case of both β_4 and β_6, we could not determine any negative parameter values that made the data simulated from M4 and M6 look different than null model simulations from model M1. We hypothesize that this is due to negative effects *removing* visually interesting structural elements as opposed to *adding* noticeable structural elements. Since we could not detect the effects, we decided that the participants in our experiment would also not be able to. So, instead of testing the negative values of these effects, we use a different scenario: we place five simulations from model M4 or M6 (with positive values of the parameter) with one simulation from model M1 in a lineup. We will refer to this as the "reverse" lineup scenario. We used the reverse scenario to determine if the perception of

[1] Please visit `https://sctyner.shinyapps.io/saom_lineup_creation/` to create lineups constructed from the models we present for this data for yourself.

Parameter	Condition	Easy Value	Medium Value	Hard Value
β_1	neg	-7.354	-6.6187	-5.883
	pos	-3.922	-4.1674	-4.412
β_2	neg	0.000	0.0005	0.049
	pos	7.340	6.8504	6.361
β_3	neg	-17.249	-10.3497	-3.450
	pos	10.350	6.8998	5.175
β_5	neg	-30.272	-20.1817	-10.091
	pos	20.182	17.6590	16.145
β_4	pos	8.351	6.6806	5.010
	reverse	6.681	5.0105	3.340
β_6	pos	5.316	3.9872	3.323
	reverse	5.316	3.9872	3.323

Table 4: All conditions used for our experiment. For parameters $\beta_1, \beta_2, \beta_3$, and β_5, M1 served as the null model. For β_4 and β_6, null model M1 and the alternative model (M4 or M6) switch roles in the reversed lineups, i.e. five plots show data simulated from the alternative model and only one plot shows data from M1.

the effect size is symmetric: if an effect is noticed $p\%$ of the time at value $\beta_k = \beta_{k_0}$ when *one* simulation from the corresponding model is placed among *five* null plots from model M1, then when *five* simulations from the model with $\beta_k = \beta_{k_0}$ are put in a lineup with *one* simulation from model M1, the plot from the simpler model should be noticed about $p\%$ of the time as well.

3.4 Execution

We recruited 250 participants for our experiment through Amazon Mechanical Turk. Each participant was presented with some brief training material first, then before presenting the lineups for the hypothesis tests, each person was shown two trial plots one where the alternative plot was the most different from the others due to its relatively *complex* structure, while the other trial included an alternative plot that was most different from the others due to its comparatively *simple* structure. Only when participants were able to correctly identify the alternative plot from the trial lineups were they allowed to begin the experiment.

Each participant was randomly assigned 13 lineups to look at. They were asked to select one or more plots that they perceived as “most different” from the others, and provide a reasoning for their choice: “most simple overall structure”,“most complex overall structure”, or “other”. If they selected “other”, they were required to describe their reasoning. The language in the reasoning is purposefully vague to avoid contextual bias.

Twelve of the 13 lineups that the participants saw were used for the significance testing and the visual power methods discussed in Sections 3.1 and 3.3. The last lineups shown to participants contained the true data from the 112th senate shown in Section 2 placed among five other plots from one of the models M3, M4, M5, and M7 as discussed in Section 3.2. Upon completion of the 13 lineups, which took an average of 308 seconds per person, each participant was paid $1.75

4 Results

In this section, we present the results from our experiment and compare them to traditional statistical tests where applicable. Because each lineup shown to participants has only six plots, the probability of picking the data by chance is high at 1 in 6, but if *many* independent viewers pick out the data from the nulls, the evidence in favor of the alternative hypothesis becomes stronger. The p-values from the lineups were calculated using the `vinference` package by Hofmann and Röttger (2016). This package contains methods to calculate *Visual distributions* for lineup experiment data. The distribution depends on the number of evaluations of a plot, L, the size of the lineup, P, and the lineup scenario, which here is that each lineup containing the same data and the same set of null plots is shown to L independent observers. The visual inference family of distributions is similar to the binomial distribution, but takes the dependency among the P plots in a single lineup shown to multiple viewers into account.

4.1 Significance Testing

For CTMC models, significance tests of the parameters are available in `RSiena`. There are t-type and Wald-type tests for a single parameter and for multiple parameters. The t-type test statistic is simply the parameter estimate divided by its standard error, and compared to a standard normal distribution (Ripley et al., 2017). The Wald-type test statistic for a single parameter, β_k is

$$\frac{(\hat{\beta}_k)^2}{var(\hat{\beta}_k)} \sim \chi^2_1, \tag{7}$$

(Ripley et al., 2017). Testing the significance of multiple parameters depends on the hypothesis we wish to test, and an $H \times K$ matrix, $\mathbf{A}$, must be appropriately designed to test the H hypotheses of interest. The null hypothesis is that $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, and the test statistic is

$$(\mathbf{A}\hat{\boldsymbol{\beta}})'\hat{\Sigma}^{-1}\mathbf{A}\hat{\boldsymbol{\beta}} \sim \chi^2_H, \tag{8}$$

where $\hat{\Sigma}$ is the estimated covariance matrix of $\boldsymbol{\beta}$.

Both parameters we test for significance using the lineup protocol, β_3 and β_4, were determined to be statistically significant using Equation 7. The corresponding results from the significance tests we performed using the lineup protocol are given in Table 5. If enough participants pick out the alternative plot to result in a p-value less than 0.05, we reject the null hypothesis that the true value of the additional parameter, either β_3 or β_4, is equal to zero.

Lineup ID	parameter	# Identified	Total Views	p-value
3131	β_3	4	29	0.60654
3132	β_3	26	31	0.00001
3133	β_3	2	27	0.80053
3141	β_4	10	23	0.03420
3142	β_4	3	37	0.77965
3143	β_4	10	29	0.09619

Table 5: Experiment results for the two parameters for which we performed significance tests. # Identified indicates the number of participants who selected the alternative model plot. There were three lineups for each parameter, so there are three results for each plot.

We see in Table 5 that the p-values for visual inference here are highly variable.

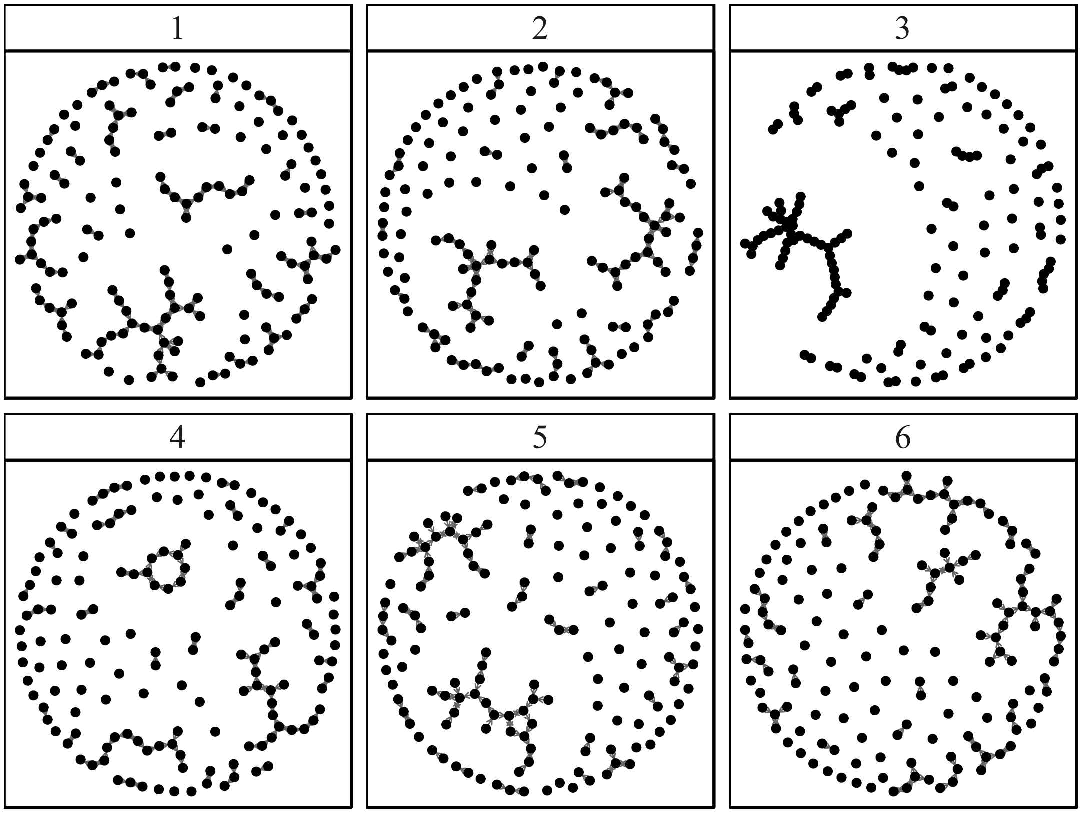


Figure 4: Lineup 3132, which led to rejection of the null hypothesis that $\beta_3 = 0$. The network simulated from model M3 is found in panel $\sqrt{16} - 1$, and the remaining panels show networks simulated from model M1.

The lineup for significance testing of β_3 which resulted in a very small p-value and rejection of the null hypothesis is shown in Figure 4. Another significance lineup for model M3, which resulted in failure to reject the null hypothesis, is shown in Figure 5. When viewing Figure 4, 26 of 31 viewers chose the alternative plot from M3, while only 2 of 27 chose the alternative plot from M3 when viewing Figure 5. The most common choice in the latter was panel two, which 16 of 27 viewers chose as the most different due to its large connected component that makes it seem more complex than the others. In viewing these two lineups, it is evident that there is a large amount of variability in networks simulated from CTMC models. It is difficult to see that five of the six networks come from the same model when they all look very different. The variability in significance test results is introduced through the null plots generated from M1. In addition, the small number of

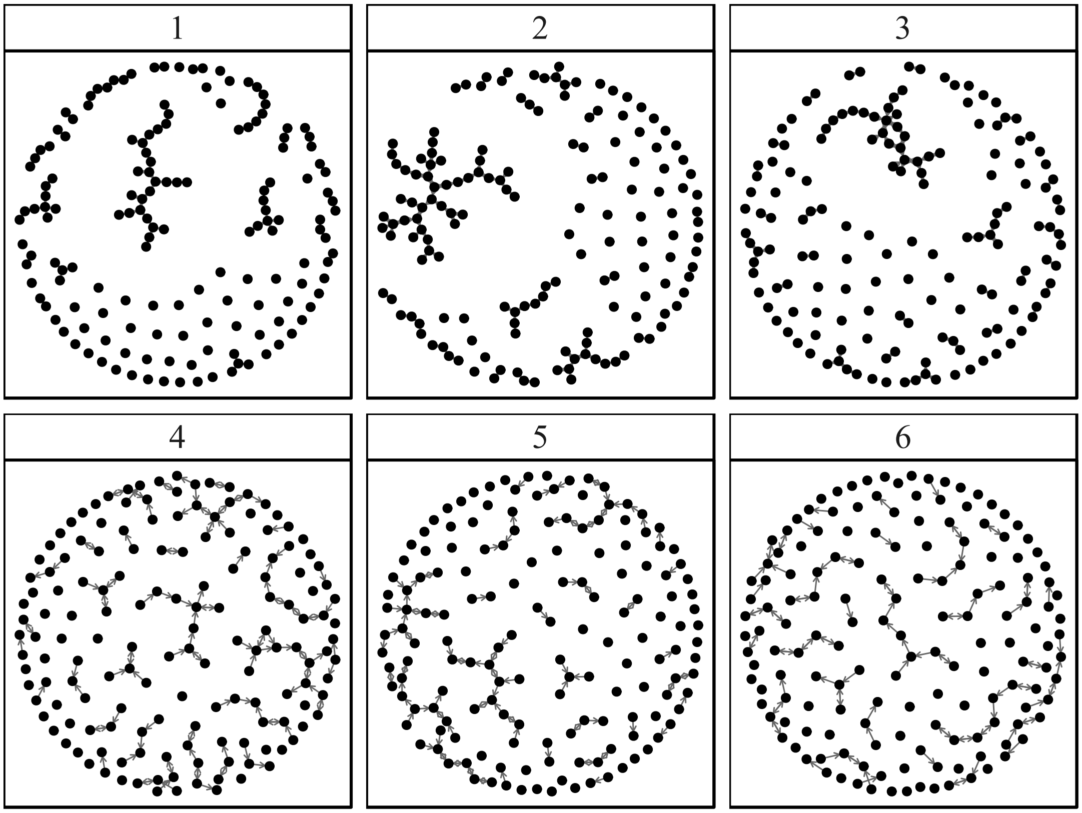


Figure 5: One of the lineups which failed to reject the null hypothesis that $\beta_3 = 0$. The network simulated from model M3 is found in panel $\sqrt{25} - 4$, and the remaining panels show networks simulated from model M1.

null plots do not give the viewer as complete of a view of the null model as the usual 19 null plots would.

The results of the significance tests given in Table 5 for β_3 and β_4 are not definitive. For the test of β_3, two of the three tests are not significant, while the third is highly significant. For the test of β_4, one test is significant, one is decidedly not significant, and the third is significant at the level of 0.10. Thus, unlike with the Wald-type tests described at the beginning of this section, we cannot decisively reject or to fail to reject the null hypothesis that the parameter value is 0.

4.2 Goodness-of-Fit Testing

Goodness-of-fit testing for network models is notoriously difficult. Most network models, other than the most simple, lack the necessary asymptotics for developing goodness-of-fit methods (Goldenberg et al., 2010). Some simulation-based methods have been developed using what Ripley et al. call *auxiliary statistics* such as the indegree or outdegree distribution on the nodes. In `RSiena`, the `sienaGOF` function performs goodness-of-fit testing as follows:

1. Auxiliary statistics, such as the cumulative outdegree counts on the nodes, are computed on the observed data ($\mathbf{u}_d$) and on N observations simulated from the model ($\mathbf{u}_1 \ldots \mathbf{u}_N$).

2. The mean $\overline{\mathbf{u}}$ and covariance matrix $\mathbf{S}$ are computed from the N simulations, and the Mahalanobis distance, $d_M(\mathbf{u}_d)$ from the observed statistics to the distribution of the simulated statistics is computed:

$$d_M(\mathbf{u}_d) = \sqrt{(\mathbf{u}_d - \overline{\mathbf{u}})'\mathbf{S}^{-1}(\mathbf{u}_d - \overline{\mathbf{u}})} \tag{9}$$

3. The Mahalanobis distance for each of the N simulations is calculated and $d_M(\mathbf{u}_d)$ is compared to this distribution of distances.

4. An empirical p-value is found by computing the proportion of simulated distances found in step 4 that are as large or larger than $d_M(\mathbf{u}_d)$.

A plot comparing the data to the simulations is also considered, and a similar plot is shown in Figure 6 for the outdegree distribution of small data set, shown in the points and connected lines, with the simulated values of $\mathbf{u}_d$ shown in box plots and overlaid violin plots. The `RSiena` software also provides a Rao score-type test for goodness-of-fit for assessing one or more parameters, the test statistic of which is compared to a Chi-square distribution with H degrees of freedom, where H has the same definition as in Section 4.1. For full detail on the score-type test, see Schweinberger (2012). These methods are both restricted: the `sienaGOF` method only considers *one* measure on the data and simulations from the model, while the score-type tests only consider *subsets* of parameters, "nuisance parameters" in Schweinberger (2012), not the entire set of parameters. By using visual inference instead of more traditional statistical methods, we hope to perform a more *holistic* goodness-of-fit test.

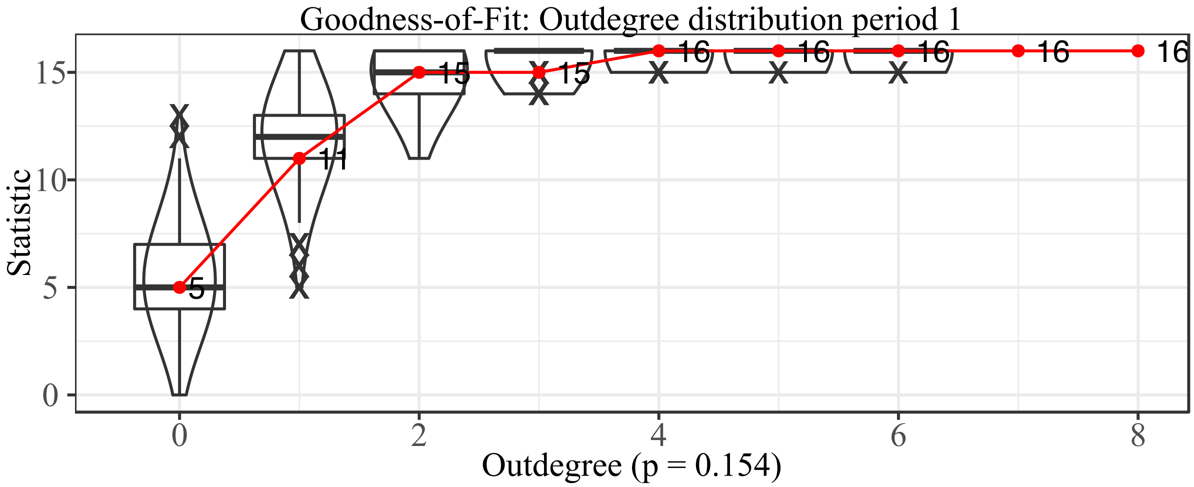


Figure 6: An example of what a goodness-of-fit plot from `RSiena` looks like. The overlaid boxplots and violin plots show the distribution of each of the outdegree count values on the simulated networks, and the red points and lines are the observed data values.

Using the lineup protocol, we show each Amazon Mechanical Turk worker the data once, in a lineup with five other plots of simulated data from one of the models we chose. We examined four different models, M3, M4, M5, and M7, and examined three repetitions of each, for a total of 12 goodness-of-fit lineups shown to participants. In each lineup, the null model is one of the four models and the alternative model is the true, unknown model that generated the senate network data. The hypotheses for our goodness-of-fit tests are:

H_0: The senate network data come from the null model, Mi.

H_A: The senate network data do not come from the null model.

If a lineup viewer picks out the data among the five simulations from the null model, it is evidence against the null hypothesis. On the contrary, if the lineup viewer picks one of the null plots, that is evidence in favor of the null hypothesis. Results from our MTurk goodness-of-fit plots are provided in Table 6.

The p-values were again computed using the `vinference` package by Hofmann and Röttger (2016). The lineup that resulted in a failure to reject the null hypothesis is shown in Figure 7. The null model in this lineup is M5, and the senate data is shown in panel number $3^2 - 7$. However, the panel most participants chose was number four, and the

Model	Replicate	Data Picks	Total Viewers	p-value
M3	1	29	36	< 0.0001
	2	13	18	0.0004
	3	16	20	< 0.0001
M4	1	13	16	< 0.0001
	2	7	20	0.1150
	3	29	34	< 0.0001
M5	1	9	21	0.0414
	2	21	24	< 0.0001
	3	14	16	< 0.0001
M7	1	17	20	< 0.0001
	2	14	28	0.0093
	3	28	37	< 0.0001

Table 6: An overview of the results from the 12 goodness-of-fit lineup tests.

most common reasoning for that choice was that it had the most simple structure. Some of the other panels, such as three and six, in Figure 7 have large connected components that are similar in size to the connected component of the data plot shown in panel two. Thus, model M5 is sometimes capable of capturing the network structure of the senate collaboration data.

The smallest *p*-value for one of the goodness-of-fit lineups was for the third replicate of the null model M5. This result contrasts with our previous finding that the only lineup to fail to reject the null was also when the null model was M5. This lineup is shown in Figure 8. In the remaining replicate of M5 as the null model, 13 of 16 viewers identified the data plot, corresponding to a *p*-values of less than 0.0001, just like the third replicate. This variability in results is similar to the variability we found in Section 4.1. This variability is again introduced through the plots simulated from null model, and does not provide us with a clear cut decision resulting the hypothesis test. For model M5, we can neither reject nor fail to reject the null hypothesis that the data come from model M5. This is evidence that the goodness-of-fit of network models cannot always be determined by one

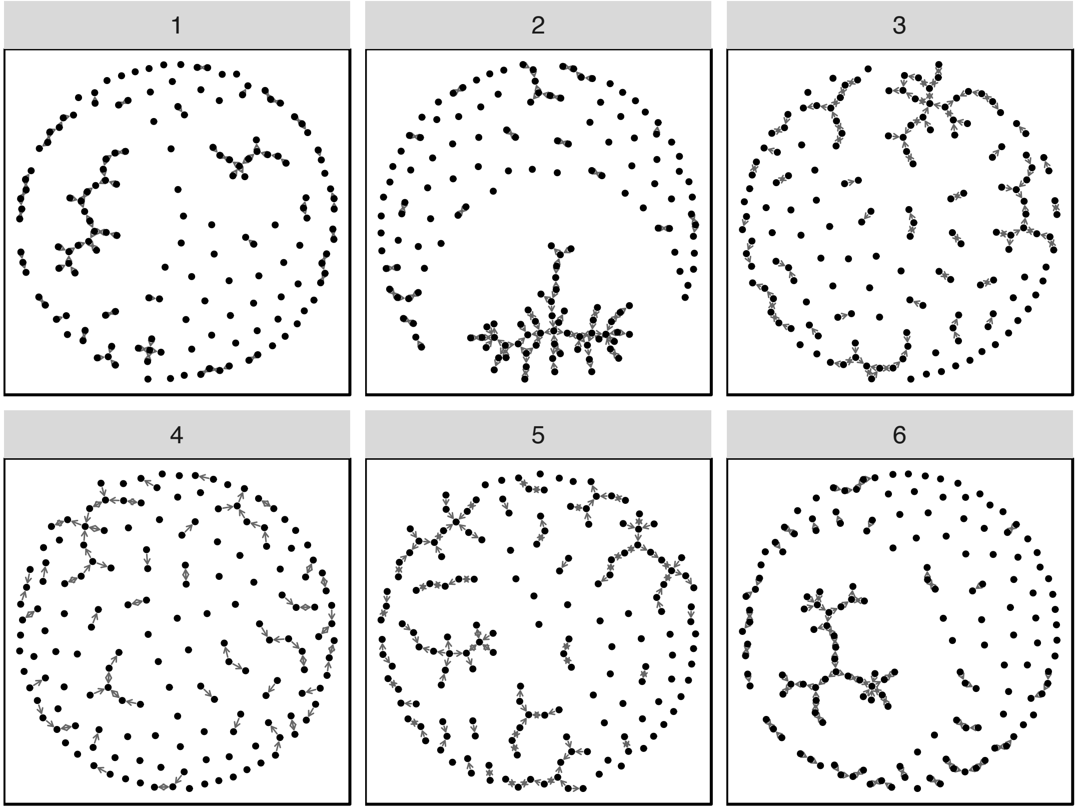


Figure 7: The goodness-of-fit lineup that failed to reject the null hypothesis. The null model for this lineup is M5. Only 7 of 20 viewers of this lineup selected the data plot as the most different from the others. The most commonly chosen panel was number four, which has a relatively simple structure compared to the other panels.

dimensional derived features.

For the other models for which we tested goodness-of-fit, however, we *do* have significant evidence from all three replicates to reject the null hypothesis that the null model generated the data. For models M3, M5, and M7, these goodness-of-fit tests have rejected the null hypotheses that the senate data come from these models. We hypothesized that the model with the most effects, M7, would be the best fit. However, as shown in Figure 9, the model does not capture the overall structure very well at all. The rest of the goodness-of-fit lineups as shown to participants are provided in the appendix.

We believe this goodness-of-fit testing method holds promise for the future of social network analysis. The participants in our experiments are very good overall at picking out the data when it is noticeably different from the null plots in the lineups. In addition, as in

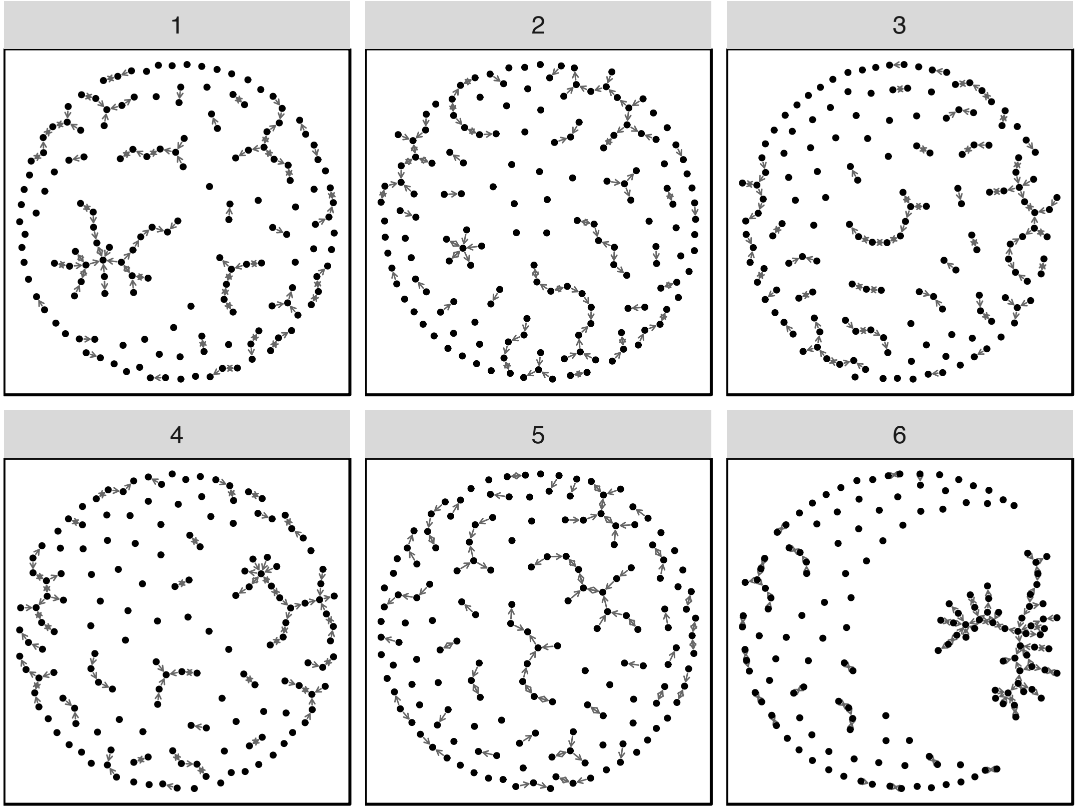


Figure 8: The lineup resulting in the smallest p-value rejecting the null hypothesis. Surprisingly, this another repetition for M5 as the null model.

replicate three for null model M4, when the null plots contain similarly sized structures as the data plot, our participants have a hard time distinguishing the data. We believe that running these tests multiple times using several different sets of null models to adequately explore the possible structures generated by the models is a step in the right direction for a more comprehensive goodness-of-fit test for network models.

4.3 Visual Power

The results from the visual power piece of our experiment are shown in Figure 10. On the x axis, we plot the value of the parameter of interest, and on the y axis, the proportion of times the alternative data plot was picked out for each lineup. The results are split into groups based on the value of the parameter and the lineup type. We can see clear patterns in the added parameters $\beta_3, \dots, \beta_6$: as the parameter value approaches 0, fewer participants identified the alternative plot. Similarly, as β_1, β_2 approach their estimated

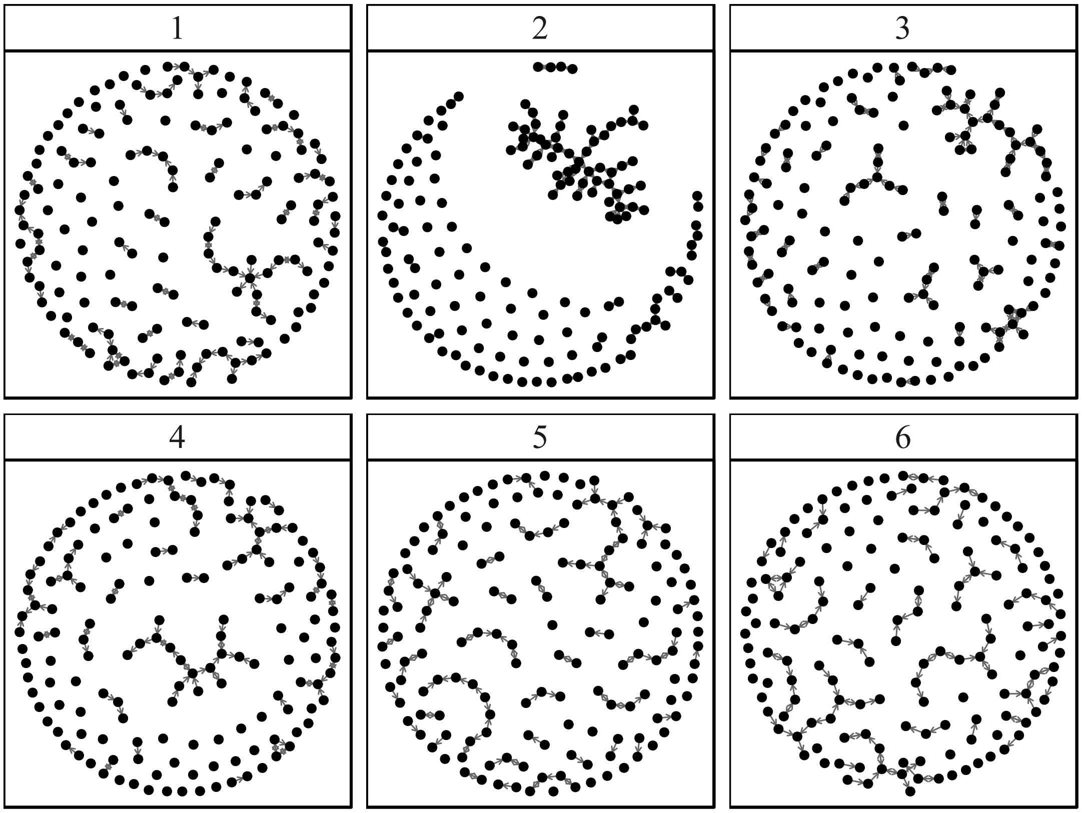


Figure 9: One repitition of a goodness-of-fit lineup testing modle M7. The senate data are shown in panel two, and it is evident that none of the other five panels, which show data simulated from model M7, come close to creating the large connected component that is central to the structure of the senate data.

values $\hat{\beta}_1, \hat{\beta}_2$, fewer people are able to identify the alternative plot.

We model the relationship between identification of the alternative data in the lineup and the parameter of interest, the effect size, and the lineup type with a generalized linear mixed model (GLMM) that provides us with an estimate of the power of the visual significance test. The response variable, X_{klqr}, is binary, indicating whether participant r picked the alternative data plot in lineup type q, rep ℓ, for effect k. There is one continuous covariate v, which is the centered and scaled size of the effect of interest from which the alternative data were simulated, the values of which are labeled "easy", "medium", and "hard" in Table 4 according to how difficult we thought the Turk participants would find each lineup. In Equation 10, $k \in \{1, 2, 3, 4, 6\}$ corresponds to the effects $\beta_1, \ldots, \beta_6$, respectively, $\ell \in \{-1, 1\}$, and $q \in \{1, 2, 3\}$. We also include random effects in the model: one for each lineup, δ_{klq}, and one for each participant, ϵ_r, and fit a hierarchical model given in Equation 10.

$$\begin{aligned} X_{klqr} &\sim \text{Bernoulli}(\pi_{klqr}) \\ \text{logit}(\pi_{klqr}) &= \eta_{kl} + \gamma_{kl} v + \delta_{klq} + \epsilon_r \\ \delta_{klq} &\overset{iid}{\sim} N(0, \sigma_\delta^2) \\ \epsilon_r &\overset{iid}{\sim} N(0, \sigma_\epsilon^2) \end{aligned} \tag{10}$$

The results of fitting this model, including estimates of parameters, standard errors, *p*-values, and the odds ratio multipliers, using `glmer` from the `lme4` package are summarized in Table 7 (Bates et al., 2015). For each combination of parameter and lineup type, the expected value of the link function for a new lineup and a new observer with parameter value v is

$$E[\text{logit}(\pi_{kl})] = \alpha_{kl} + \gamma_{kl} v \tag{11}$$

and the corresponding probability of picking out the alternative data plot is

$$\pi_{kl} = \frac{\exp\{\alpha_{kl} + \gamma_{kl} v\}}{1 + \exp\{\alpha_{kl} + \gamma_{kl} x\}} \tag{12}$$

In Figure 10, we see a clear trend in all parameters except β_1 and β_2 that as the parameter value approaches zero from either side, the probability of picking the data plot in a lineup of size six decreases. This supports our hypothesis shown in Figure 3. For β_3 and β_5, the slope of the fitted line is *much* steeper for positive values of the parameter than for negative values, meaning that our participants perceived differences more often for positive parameter values than for negative parameter values. This finding is similar to that of Harrison et al. (2014), who found that people detect positive correlations better and at lower values than negative correlations.

We expand portions of Figure 10 in Figures 11-13. These figures show the same prediction regions as in Figure 10, plus some additional predictions outside of the data range shown in light gray. Again, the points represent the results from the experiment. In all three of these figures, the lack of symmetry is apparent. In the reverse lineup scenario shown in Figure 13, the probability of prediction is consistently far less than the probability of prediction in the regular lineup scenario. This demonstrates that the visual signal of

Parameter	Estimate	Std Error	p-value	Odds Multiplier
η_{1+}	37.297	6.573	<0.0001‡	>1e+06
η_{1-}	-9.933	3.473	0.0042†	<0.0001
γ_{1+}	75.356	13.532	<0.0001‡	>1e+06
γ_{1-}	-14.504	4.804	0.0025†	<0.0001
η_{2+}	-6.833	4.466	0.1260	0.0011
η_{2-}	-17.001	2.236	<0.0001‡	<0.0001
γ_{2+}	13.771	7.446	0.0644*	956752.4844
γ_{2-}	-229.16	31.306	<0.0001‡	<0.0001
η_{3+}	-2.801	0.949	0.0032†	0.0608
η_{3-}	-2.644	0.811	0.0011†	0.0711
γ_{3+}	4.474	1.389	0.0013†	87.7507
γ_{3-}	-1.108	0.609	0.0690*	0.3304
η_{4+}	-2.078	0.954	0.0293**	0.1252
η_{4-}	-2.692	1.322	0.0417**	0.0678
γ_{4+}	4.247	2.147	0.0479**	69.8675
γ_{4-}	2.403	2.187	0.2719	11.0585
η_{5+}	-5.84	2.989	0.0507*	0.0029
η_{5-}	-4.686	0.86	<0.0001‡	0.0092
γ_{5+}	3.264	1.756	0.0630*	26.1487
γ_{5-}	-2.176	0.387	<0.0001‡	0.1136
η_{6+}	-1.164	1.226	0.3425	0.3123
η_{6-}	-5.929	1.28	<0.0001‡	0.0027
γ_{6+}	5.76	3.524	0.1021	317.4753
γ_{6-}	15.092	3.605	<0.0001‡	>1e+06
σ^2_{δ}	0.564	–	–	–
σ^2_{ϵ}	0.342	–	–	–

Table 7: Summary of the results from fitting the model given in Equation 10. Significance levels: * - < 0.10; ** - < 0.05; † - < 0.01; ‡ - < 0.001

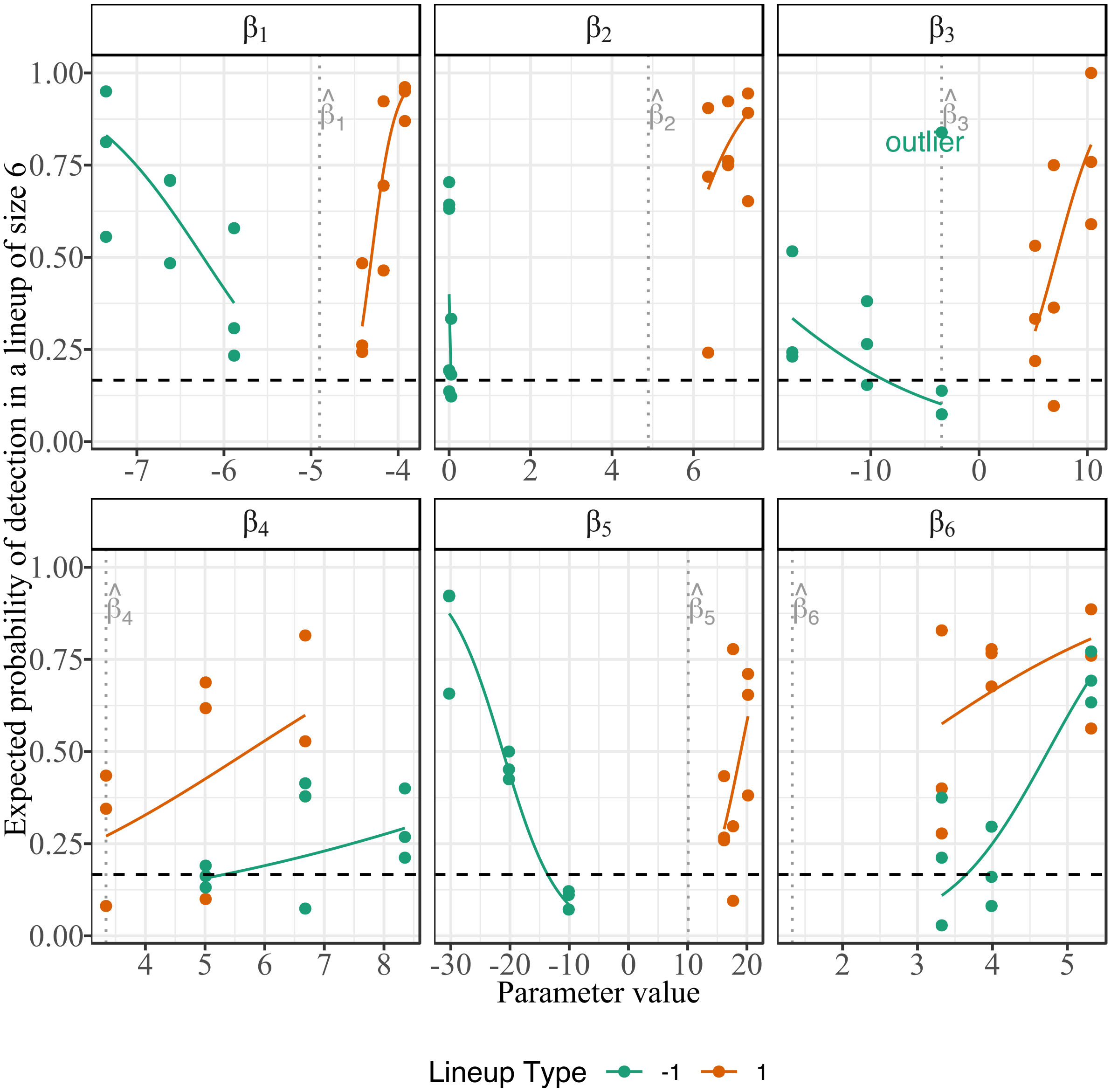


Figure 10: Predictions from our GLMM given in Equation 10. For new observers of new lineups, the lines show the expected probability of detecting the alternative data in a lineup of size 6 as a function of the parameter value. The proportions detected by our Turk participants for each lineup group are shown by the points, with the probability of picking out the data plot at random shown by a horizontal line at 1/6. The lineup marked as "outlier" was removed from modeling. Estimated parameter values shown by vertical dotted lines.

one plot from M4 among five plots from M1 is much stronger than that of one plot from M1 among five plots from M4. We posit that the latter is a more difficult task because it

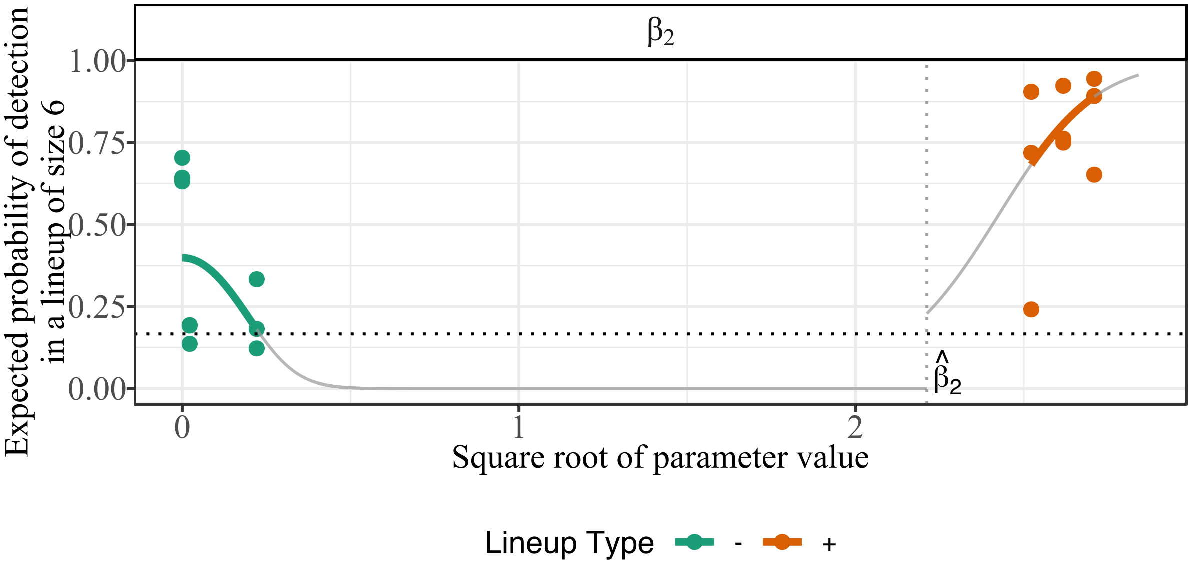


Figure 11: The top middle panel of Figure 10 expanded to show greater detail. The square root of the parameter value is shown on the x-axis. For this parameter, as its value approaches zero, the probability of identifying the alternate data model decreases, then increases, which is noticeably different from the pattern exhibited by the others. Again, a horizontal line is drawn at 1/6, the chance of selecting the data plot at random.

involves noticing a *lack of structure* as opposed to the presence of *more* structure. We can see a similar effect in Figure 12. At a value of $\beta_5 = 20$, the model predicts a probability of about 0.60 that a new viewer of a new lineup will identify the alternative data plot. At a value of $\beta_5 = -20$, however, the model predicts this same probability to be about 0.40. This again demonstrates that the *presence* of structure is detected more frequently and at smaller values than the absence of structure.

For β_4 and β_6, where one plot simulated from M1 was placed among five plots from the corresponding model, we see that the predictions for the reverse lineup type (-1), are less than the standard lineup type (1) for all values of the parameter that we have. This contradicts our hypothesis for this scenario, which was that these two scenarios would perform similarly. One of the lineups for the $\beta_4 = 6.681$, lineup type 1 scenario is given in Figure 14, and a corresponding lineup for the lineup type -1 scenario is given in Figure 15. For identical values of the parameter, viewers had a harder time identifying the different plot when they were selecting the most "simple" structure, detecting M1 in five plots from

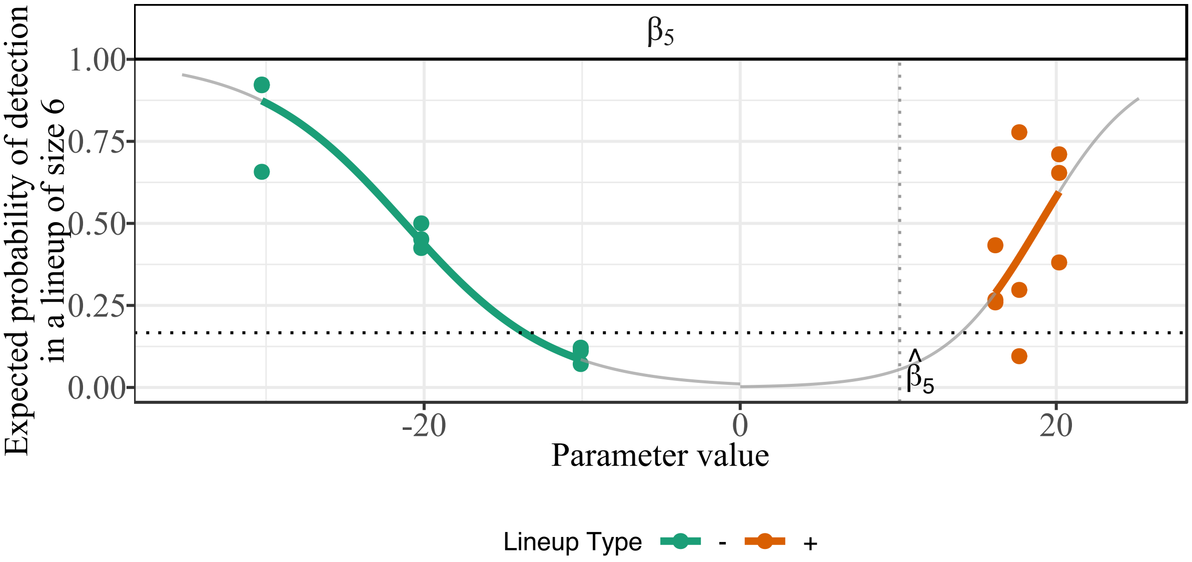


Figure 12: The bottom middle panel of Figure 10 expanded to show greater detail. The parameter value is shown on the x-axis. This parameter most closely follows our hypothesis shown in Figure 3. However, the result is not symmetric. According to the model, people will detect the effect at lower values and with greater frequency as the value increases when it is positive instead of negative.

the more complicated model, than they did identifying the most "complex" structure, the plot from the more complicated model, from the five plots from M1.

5 Discussion

By using visual inference methods, we have developed new ways to perform significance and goodness-of-fit tests for a complicated, intractable set of statistical models for social network data. We have also developed a way to determine the power of these new visual tests. Our methods can be used to supplement traditional methods and check our assumptions about network models. The traditional methods only look at one piece or derived measure of a network model, whereas our methods look at the models holistically for a broader sense of what it means for a parameter to be significant or a model to be a good fit. By looking at an entire network simulated from a network model side-by-side with other instances of networks simulated from a null model, instead of singular features, we develop an idea of

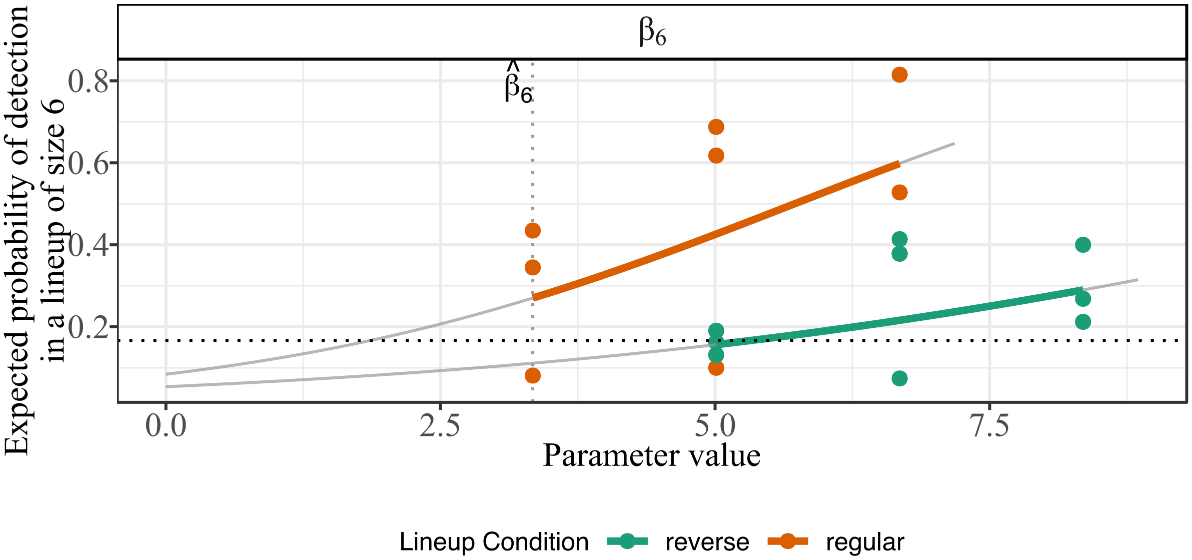


Figure 13: The bottom right panel of Figure 10 expanded to show greater detail. The parameter value is shown on the x-axis. The "reverse" lineup has a much flatter slope than the "regular" lineup, which means the participants had a harder time detecting a more simple M1 structure among many more complex M4 structures. Reversing the lineup scenario was not symmetric as we hypothesized.

the model in terms of the *data* itself, not in terms of statistical summaries of the data. These methods place the model in the data space: they don't summarize or compress the data to put it in the model space (Wickham et al., 2015).

Furthermore, we have found the visual power of some effects in the object function of a CTMC model for this particular senate data example, and we have shown that, for the same effects, there is a lot of variability in results from significance and goodness of fit tests. Because the visual tests we performed show a great deal of variability, we can see that the decisions with respect to the significance of a parameter or the goodness of fit of a model to data are not as cut-and-dried as the more traditional methods would have us believe.

These results do not come without limitations. In VI, the null plots are supposed to play the role of good representatives of the null model. Here, the number of null plots is reduced to five, which has lead to very different conclusions for the same lineup scenario. Furthermore, these results do not generalize to all CTMC models or to one particular

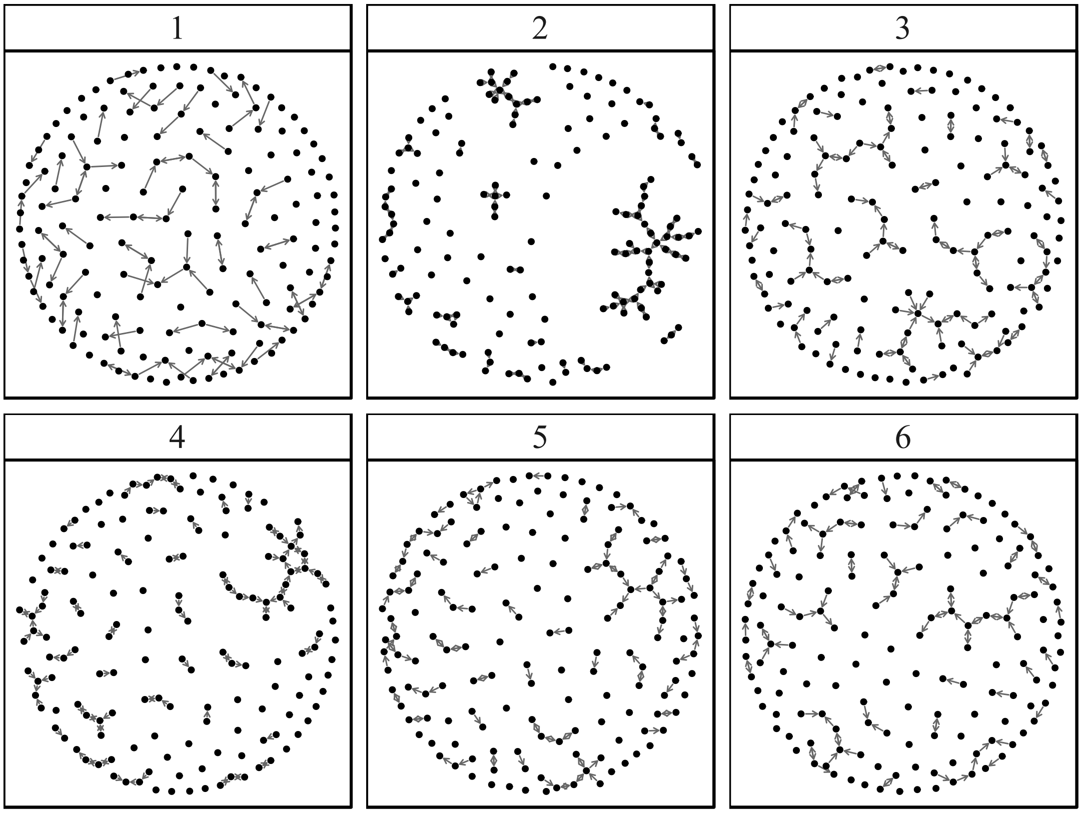


Figure 14: In our experiment, 52.8% of viewers of this plot selected the plot from the alternative model, M4. The "reverse" of this lineup is given in Figure 15, where 41.4% of viewers selected the plot from the alternative model, M1. Here, the alternative plot is $\sqrt{25} - 3$.

subset of CTMC models. The lineups shown are made for only one set of data, and it is not clear that the visual power results will transfer to other situations with different number of actors, different edge densities, or different layout algorithm of the node-link diagram. We can make some generalizations about what participants are picking up on in the lineups based on their feedback and previous research, but we cannot apply our hierarchical model directly to lineups constructed for new data or new models or parameters.

We these methods will be applied to different types of network data and different types of network models. But given the limitations of the node-link visualization, the cognitive load of looking at a lineup is very high for the average observer. More research is needed to apply these methods to larger datasets, different layout algorithms, and different ways of visualizing network data, such as adjacency matrix visualizations, using visual inference to see if similar findings emerge.

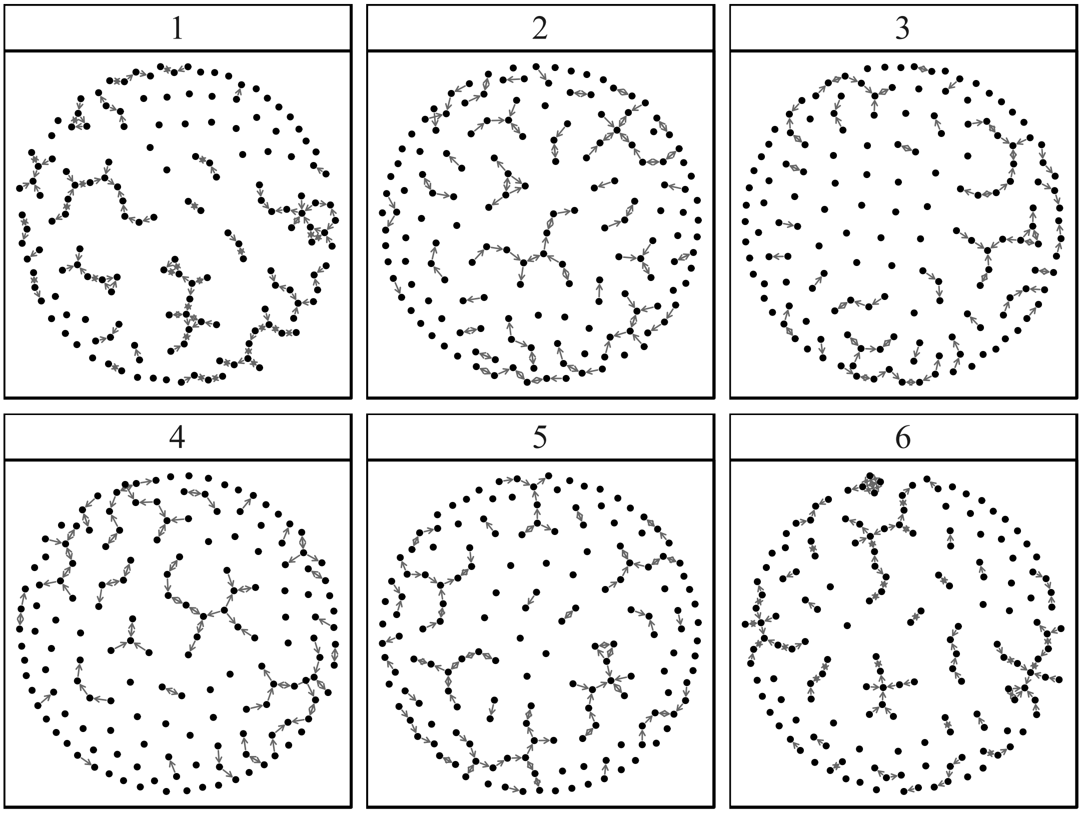


Figure 15: In our experiment, 41.4% of viewers of this plot selected the plot from the alternative model, M1. The “reverse” of this lineup is given in Figure 14, where 52.8% of viewers selected the plot from the alternative model, M1. Here, the alternative plot is $\sqrt{25} - 1$.

References

Amazon (2010).

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), “Fitting Linear Mixed-Effects Models Using lme4,” *Journal of Statistical Software*, 67, 1–48.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E., Swayne, D., and Wickham, H. (2009), “Statistical Inference for Exploratory Data Analysis and Model Diagnostics,” *Royal Society Philosophical Transactions A*, 367, 4361–4383.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2017), *shiny: Web Application Framework for R*, r package version 1.0.3.

Fruchterman, T. M. and Reingold, E. M. (1991), "Graph Drawing by Force-Directed Placement," *Software: Practice and Experience*, 21, 1129–1164.

Gelman, A. (2004), "Exploratory Data Analysis for Complex Models," *Journal of Computational and Graphical Statistics*, 13, 755–779.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. (2010), "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning*, 2, 129–233.

Gross, J. H., Kirkland, J. H., and Shalizi, C. R. (2008), "Cosponsorship in the U.S. Senate: A Multilevel Two-Mode Approach to Detecting Subtle Social Predictors of Legislative Support," *Unpublished Manuscript.*

Harrison, L., Yan, F., Franconeri, S., and Chang, R. (2014), "Ranking Visualizations of Correlation Using Weber's Law," *IEEE Transactions on Visualization and Computer Graphics*, 20, 1943–1952.

Hofmann, H. and Röttger, C. (2016), *vinference: Inference under the lineup protocol*, r package version 0.1.1.

Holland, P. and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs (with discussion)," *Journal of the American Statistical Association*, 76, 33–65.

Hummel, R. M., Hunter, D. R., and Handcock, M. S. (2012), "Improving Simulation-Based Algorithms for Fitting ERGMs," *Journal of Computational and Graphical Statistics*, 21, 920–939.

Loy, A., Follett, L., and Hofmann, H. (2015), "Variations of Q-Q Plots – the Power of our Eyes!" *The American Statistician*, 2015, 1–36.

Majumder, M., Hofmann, H., and Cook, D. (2013), "Validation of Visual Statistical Inference, Applied to Linear Models," *Journal of American Statistical Association*, 108, 942–956.

Ringe, N., Victor, J. N., and Cho, W. T. (2016), *The Oxford Handbook of Political Networks*, Oxford University Press, chap. Legislative Networks.

Ripley, R., Boitmanis, K., and Snijders, T. A. (2013), *RSiena: Siena - Simulation Investigation for Empirical Network Analysis*, r packa ge version 1.1-232.

Ripley, R. M., Snijders, T. A., Boda, Z., Vörös, A., and Preciado, P. (2017), "Manual for RSiena," Tech. rep., `https://www.stats.ox.ac.uk/~snijders/siena/RSiena_Manual.pdf`.

Schweinberger, M. (2012), "Statistical modelling of network panel data: Goodness of fit," *British Journal of Mathematical and Statistical Psychology*, 65, 262–281.

Snijders, T., Steglich, C., and Schweinberger, M. (2007), *Longitudinal Models in the Behavioral and Related Sciences*, Lawrence Erlbaum Associates, chap. Modeling the Coevolution of Networks and Behavior.

Snijders, T. A. (2001), "The Statistical Evaluation of Social Network Dynamics," *Sociological Methodology*, 31, 361–395.

— (2005), *Models and Methods in Social Network Analysis*, New York: Cambridge University Press, chap. Models for Longitudinal Network Data.

— (2017), "Stochastic Actor-Oriented Models for Network Dynamics," *Annual Review of Statistics and Its Application*, 4, 343–63.

Snijders, T. A., van de Bunt, G. G., and Steglich, C. E. (2010a), "Introduction to stochastic actor-based models for network dynamics," *Social Networks*, 32, 44 – 60, dynamics of Social Networks.

Snijders, T. A. B. (1996), "Stochastic actor-oriented models for network change," *Journal of Mathematical Sociology*, 21, 149–172.

Snijders, T. A. B., Koskinen, J., and Schweinberger, M. (2010b), "Maximum likelihood estimation for social network dynamics," *The Annals of Applied Statistics*, 4, 567–588.

Swan, M. (2013), "The Quantified Self: Fundamental Disruption in Big Data Science and Biological Discovery," *Big Data*, 1, 85–99.

Tyner, S. and Hofmann, H. (2018), *geomnet: Network Visualization in the 'ggplot2' Framework*, r package version 0.3.0.

Vander Plas, S. and Hofmann, H. (2015), "Clusters beat Trend!? Testing feature hierarchy in statistical graphics," *Journal of Computational and Graphical Statistics*, submitted.

Wickham, H., Cook, D., and Hofmann, H. (2015), "Visualizing statistical models: Removing the blindfold," *Statistical Analysis and Data Mining*, 8, 203–225.