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

August 6, 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 visual inference can be used to supplement traditional statistical tests for network models.

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

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, people are *nodes* or *actors* connected by relationships, which are *edges* or *ties*. 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, 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* to complement

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, which can lead to very complicated models. The model complexity and the inherent complexity of network data combine to make interpreting model parameters and their estimates veryh difficult. citation? 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. The model is hierarchical: first, the *rate function* dictates *how often*

3

changes in the network occur, then the *objective function* determines *what* those changes are.

Rate Function: The rate function dictates when network changes are made and which actor can make them. An actor, i, is chosen make a change in its ties 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 which directly impact the overall network structure. We denote the rate from t_m to t_{m+1} as α_m for $m = 1, \ldots, M-1$, where M is the number of time points at which the network was observed. 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 is selected to make a change, it randomly picks one of its current ties, x_{ij}, to change. 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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ are additional model parameters, each associated with some statistics, $s_{i1}(x, \mathbf{Z}), \ldots, s_{iK}(x, \mathbf{Z})$, calculated for actor i at the current network state x. At 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 $s_{i1}(x) = \sum_j x_{ij}$ 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 $f_i(x, \boldsymbol{\beta}, \mathbf{Z})$ dictates 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. Thus, the actor is more likely to make changes that increase the value of their objective function, and no change is most likely when any change decreases the value of the objective function.

1.2 Visual Inference

Data visualization is 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 for the CTMC model. Under the lineup 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 the model of interest. 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, though an observer has a chance of 1 in P to pick the data plot at random. 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 111th through 114th Congresses. These senates began on January 6, 2009, the start date of the 111th Congress, and ended on January 3, 2017, the last date of the 114th Congress. 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.

Models: We fit six models to the senate data, and each model is identified by the parameters its objective function as shown in Table 2. 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-type as described in Section 4 to perform significance tests on $\beta_3, \ldots, \beta_6$ in models M3,$\ldots$, 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.
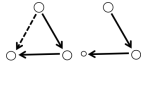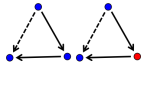
β_k	Effect name	Interaction Variable	Formula	Picture	Initial estimate	Wald p-value
β_1	density	–	$s_{i1}(x) = \sum_j x_{ij}$		2.204	NA
β_2	reciprocity	–	$s_{i2}(x) = \sum_j x_{ij} x_{ji}$		-4.903	NA
β_3	jumping transitive triplet	party	$s_{i3}(x, \mathbf{p}) = \sum_{j \neq h} x_{ij} x_{ih} x_{hj} \cdot \mathbb{I}(p_i = p_h \neq p_j)$		-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	$s_{i5}(x, \mathbf{b}) = \sum_j x_{ij} x_{ih} x_{hj} \cdot (sim_{ij}^b - \overline{sim}^b)^*$		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 for 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_{ij}^b is defined in Equation 3 and $\overline{sim}^b = \frac{1}{n(n-1)} \sum_{i \neq j} sim_{ij}^b$ 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.
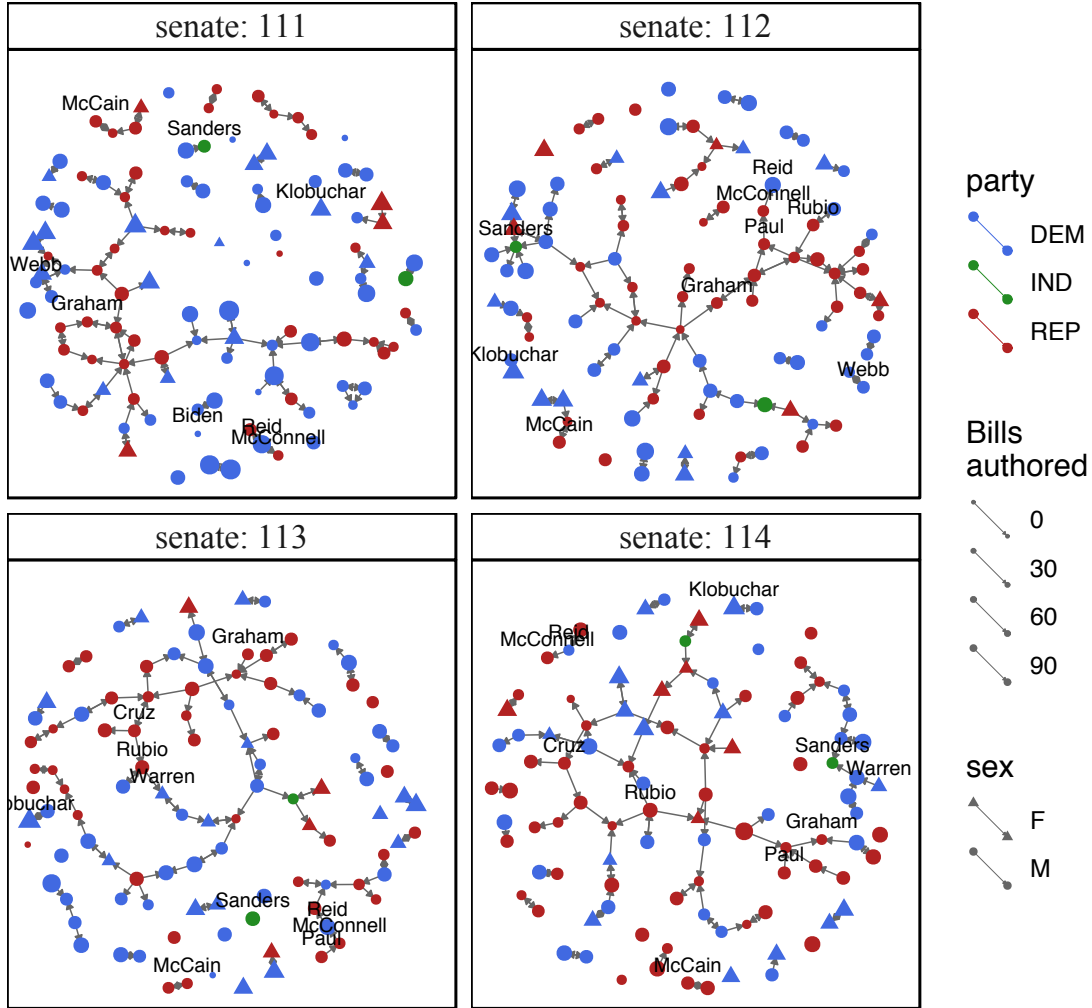
Figure 1: The US senate collaboration network observed at four time points. Color represents party, shape represents gender, and size represents number of bills authored in a session. The Fruchterman-Reingold layout is shown (Fruchterman and Reingold, 1991). Drawn with the `geomnet` R package. (Tyner and Hofmann, 2018)

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 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

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 mean estimates from repeated fits of our models of interest. When simulating from these models, these are the estimates that we will use unless otherwise stated.

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:

$$sim_{ij}^b = \frac{\max_{hk} |b_h - b_k| - |b_i - b_j|}{\max_{hk} |b_h - b_k|} \tag{3}$$

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. That senators tend co-sponsor bills written by senators who are similarly prolific corresponds to the tendency of senators to be either "workhorses" or "show horses" (Ringe et al., 2016). 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 that 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. We use the estimates given in Table 3 to simulate from models M1 through M6. Then, we examine the effects on the structure of the simulated networks with a large visual inference experiment to determine significant effects, goodness-of-fit, and visual power.

3   Methodology

We want to assess CTMC models for the senate network data using the lineup protocol in three ways: with significance tests of parameters, goodness-of-fit tests of a model, and determination of the 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. 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 112th Senate network, and we condition on the first wave of data, 111th Senate network. Sections 3.1 through 3.3 describe in detail how we constructed the lineups, which parameter values used, 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.

We recruited 250 participants for our experiment through Amazon Mechanical Turk. Each participant was first presented with some brief training material. Before presenting the lineups for the hypothesis tests, each participant 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. Among the 12 lineups were four from each difficulty level (easy, medium, and hard), and two of the four were from each condition (negative-positive, or positive-reverse). Each participant saw each of the six models twice at some combination of difficulty and condition. 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, each participant was paid $1.75

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:

$$H_0 : \beta_k = 0 \quad \text{versus} \quad H_A : \beta_k \neq 0 \tag{4}$$

Under the null hypothesis, we assume that the model that generated the network data is M1, the simplest model presented in Section 2, and the alternative model is the one with β_k

included. 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 data plot is simulated from the appropriate model with β_1, β_2, and β_k in the objective function at corresponding values given in Table 3. We constructed the lineups in this way because we hypothesize that if an effect is significant, it will alter the appearance of the simulated networks to the extent that they can be visually distinguished from the networks simulated from the simple model M1. Thus, an observer picking the alternative data plot as most different 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. We compare the null model M1 to the alternative models M3 and M4, using three repetitions of each hypothesis test in the experiment to determine if β_3, β_4 are significant.

3.2   Goodness-of-Fit Testing

For the goodness-of-fit tests, we compare one model of interest to the second wave of sentate network data. This procedure is identical to that in Buja et al. (2009) and Majumder et al. (2013). The hypotheses for the goodness-of-fit tests 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 second wave networks from the model of interest using the corresponding parameters in Table 3. Among these five plots, we place the true second wave of data. We cannot show the data more than once to each participant because doing so would bias the results, so each participant only sees one goodness-of-fit lineup. The models we chose for goodness-of-fit testing are M3, M4, M5, and M7. If a participant selects the data as most different, that is evidence that the model of interest is not a good fit.

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 plot containing the 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, $\beta_1, \ldots, \beta_6$.

In model M1, with only two parameters in the objective function, we changed the density and reciprocity parameter values one at a time, keeping all other parameters constant. We look at some values greater and some values lower than the estimates of β_1, β_2 given in Table 3. In models M3 through M6, we vary only the additional parameter, β_3 through β_6, respectively, while holding all other parameters constant. 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.
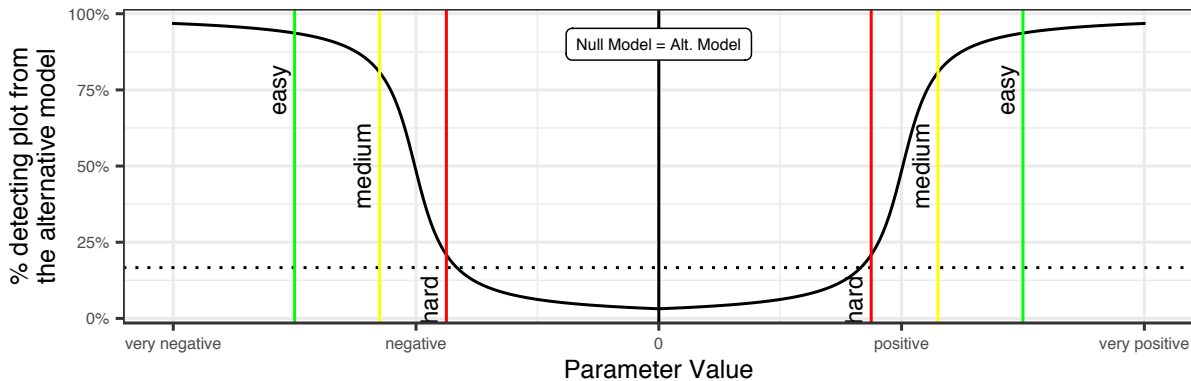

Figure 2: As the parameter increases in absolute value, more viewers of the lineup should pick the alternative data out of a lineup. Note that the significance test we construct in Section 3.1 is just one point on the curve. The easy, medium, and hard lines represent how we determined which values of the parameters to show to our participants, and the horizontal dotted line is the chance of picking the alternative plot at random.

To determine when an effect becomes noticeable, we examine six different levels of the effect: three negative and three positive. Figure 2 shows a sketch of the hypothetical selec-

Figure 3: A screen shot of the web application we created to help design our lineup experiment.

tion probability with varying effect size. We hypothesize that as the parameter increases in absolute value, an observer is more likely to select the alternative data plot out of the lineup. The six levels of the effect are vertical lines labeled "easy," "medium," and "hard" in Figure 2. We expect most observers to pick out the alternative data plot at the "easy" values, and we expect very few, if any, observers to pick out the alternative data plot at the "hard" values. To help us decide on the parameter values to use for each effect at 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 3. On the left side of the screen, the user[1] can input the information necessary for creating a lineup using network data simulated from models M1 through M7. The plots in the lineup that are not specified by the user contain data simulated from model M1.

Using the Shiny application, we settled on six parameter values to test for each of the six effects, $\beta_1, \ldots \beta_6$. 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

---

[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	6.681	5.0105	3.340
	reverse	8.351	6.6806	5.010
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.

made the networks simulated from M4 and M6 look different than the networks simulated 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 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.

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 against the null 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. Suppose that out of L observers of a lineup of size P, there are ℓ observers who pick out the alternative data plot. Then the corresponding p-value from the visual distribution gives the probability that ℓ or more out of L independent observers would pick out the alternative data plot by chance.

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

$$\frac{(\hat{\beta}_k)^2}{var(\hat{\beta}_k)} \sim \chi^2_1, \tag{5}$$

(Ripley et al., 2017). Both parameters we test for significance using the lineup protocol, β_3 and β_4, were determined to be statistically significant using Equation 5 with p-values given in Table 5. For the visual significance test, if the number of participants who pick

out the alternative plot results 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. If the null hypothesis is not rejected, then there is evidence that the additional parameter does not affect the overall structure of the network even though it was found to be sigificant with the Wald-type test. The p-values from the visual distribution for the significance tests of β_3, β_4 are given in Table 5. The results for each lineup, whether to reject or fail to reject the null hypothesis, vary, with one of three lineups from each parameter group resulting in rejection of the null hypothesis.

Lineup ID	parameter	# ID	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. # ID indicates the number of participants who identified the alternative data plot. There were three lineups for each parameter, so there are three results for each plot.

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. The variability in significance test results is introduced through the null plots generated from M1. It is difficult to see that five of the six networks come from the same model when they all look very different. In addition, the small number of null
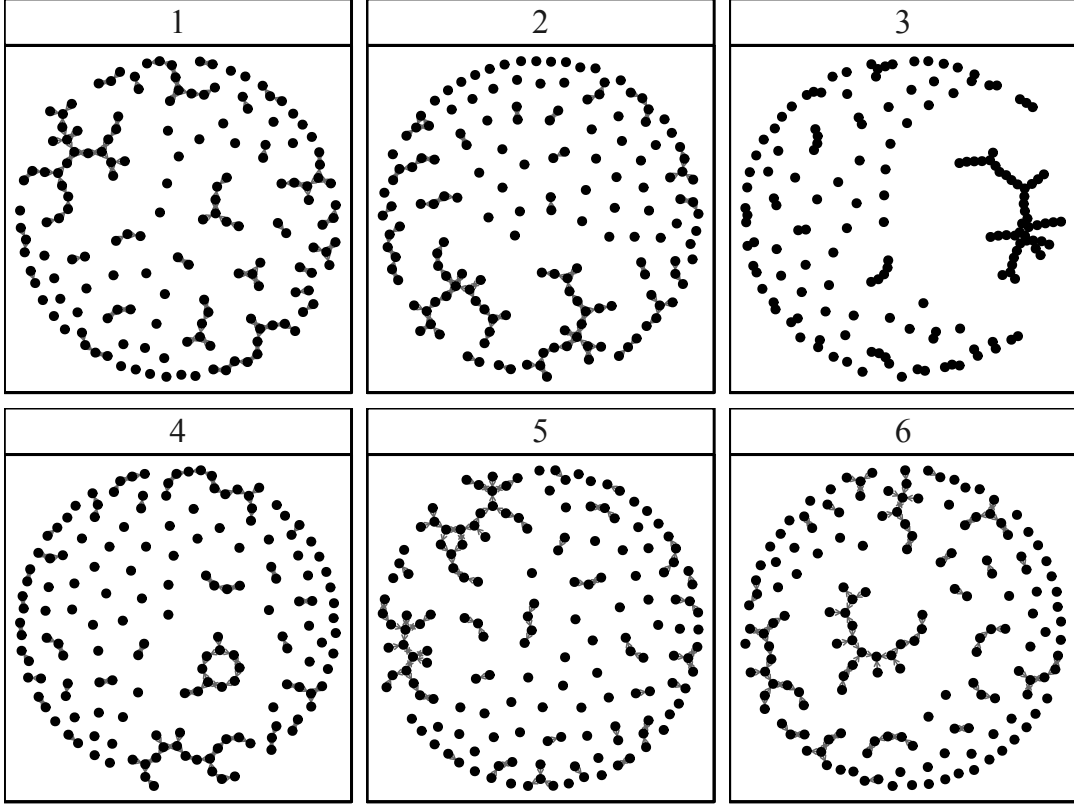
Figure 4: Lineup 3132, which led to rejection of the null hypothesis that $\beta_3 = 0$. The network simulated from model M3 is found in panel $\sqrt{16} - 1$, and the remaining panels show networks simulated from model M1.

plots do not give the viewer as complete of a view of the null model as the usual 19 null plots would. In both lineups in Figures 4 and 5, most participants chose the plot with the largest connected component as the most different.

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.
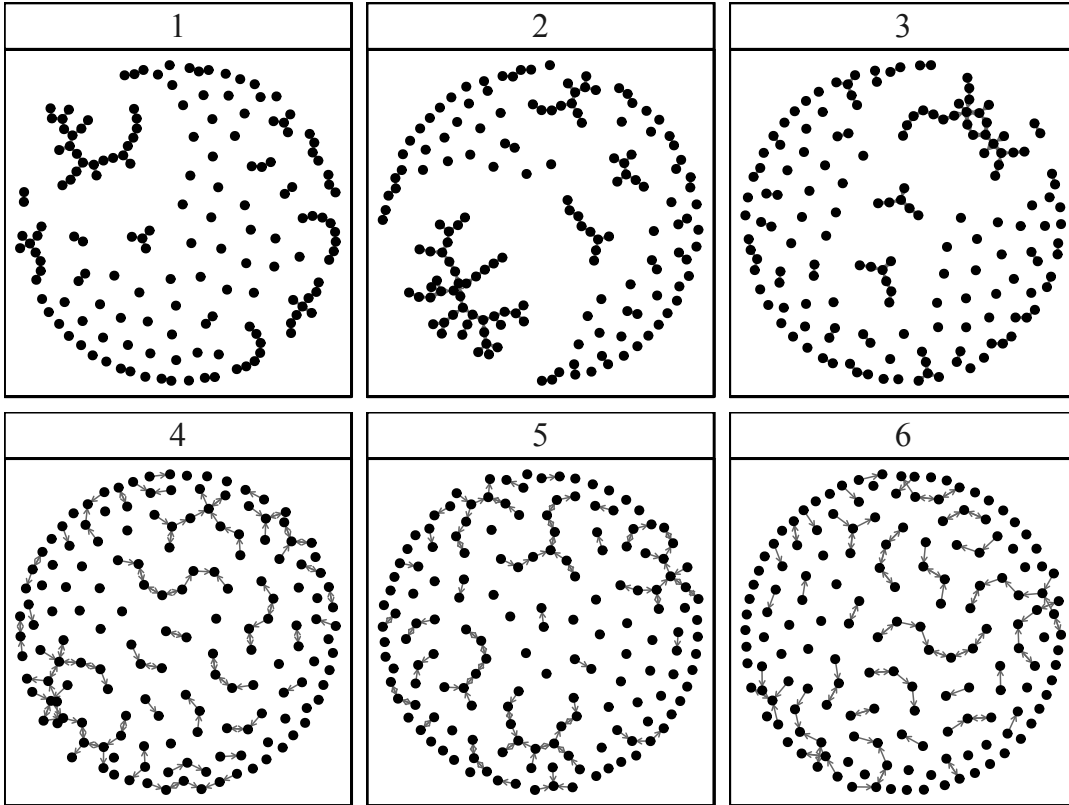
Figure 5: Lineup 3133, which resulted in failure to reject the null hypothesis that $\beta_3 = 0$. The network simulated from model M3 is found in panel $\sqrt{25} - 4$, and the remaining panels show networks simulated from model M1. 16 of 27 viewers picked plot two as most different.

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 auxiliary statistics such as the indegree or outdegree distribution on the nodes (Ripley et al., 2017). In `RSiena`, @ This goodness-of-fit test is very limited because it only considers one measure on the data and simulations at a time. To assess goodness-of-fit on the whole network, many auxiliary statistics need to be considered. Thus, by using visual inference instead of more traditional statistical methods, we construct a more holistic goodness-of-fit test.

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. 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.

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.

The p-values were again computed using the `vinference` package by Hofmann and Röttger (2016). In all lineups except replicate 2 of M4, the null hypothesis that the data were generated by the model is rejected. Thus, none of the models we have chosen are a good fit to the data. The lineup that resulted in a failure to reject the null hypothesis is shown in Figure 6. The null model in this lineup is M4, and the senate data is shown in panel number $3^2 - 7$. However, the panel most participants chose was number four, and the 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 6 have large connected components

that are similar in size to the connected component of the data plot shown in panel two. Thus, model M4 is sometimes capable of capturing the network structure of the senate collaboration data.
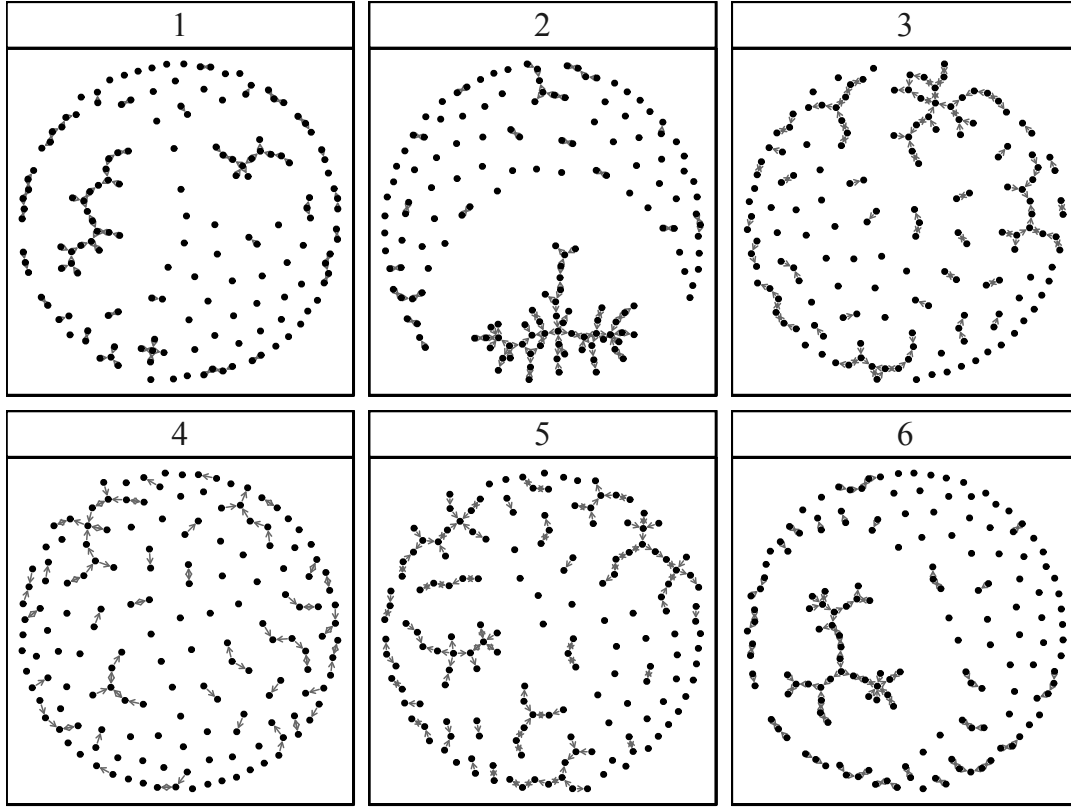

Figure 6: The goodness-of-fit lineup which resulted in failure to reject the null hypothesis. The null model for this lineup is M4. Only 7 of 20 viewers of this lineup selected the data plot as the most different from the others. The most commonly chosen panel was number four, which has a relatively simple structure compared to the other panels.

The smallest p-value for one of the goodness-of-fit lineups was for the third replicate of the null model M4. This result contrasts with our previous finding that the only lineup to fail to reject the null was also when the null model was M4. This lineup is shown in Figure 7. In the remaining replicate of M4 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, and does not provide us with a clear cut result from the goodness-of-fit tests. For model M4, we can neither reject nor fail to reject the null hypothesis that the data come from model M4.

21

This is evidence that the goodness-of-fit of network models cannot always be determined by one dimensional derived features.
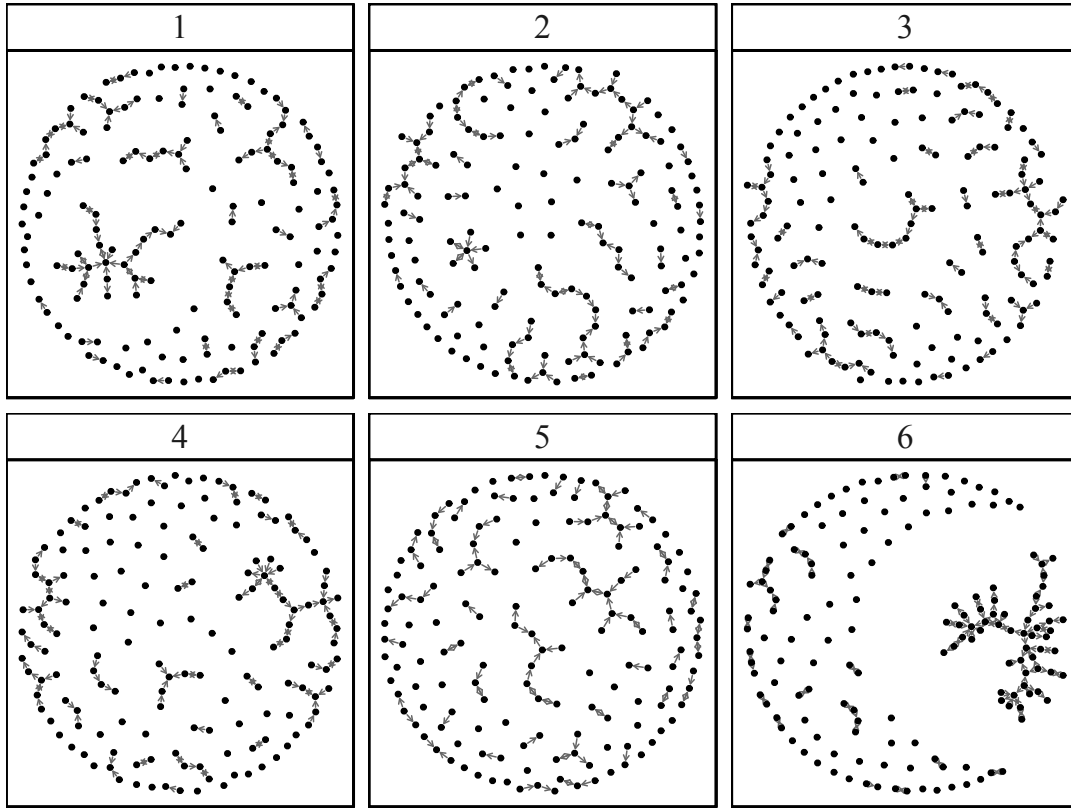

Figure 7: The lineup resulting in the smallest p-value rejecting the null hypothesis. Surprisingly, this another repetition for M5 as the null model. The data are in panel 6.

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 8, 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 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 replicate three for null model M4, when the null plots contain similarly sized structures as
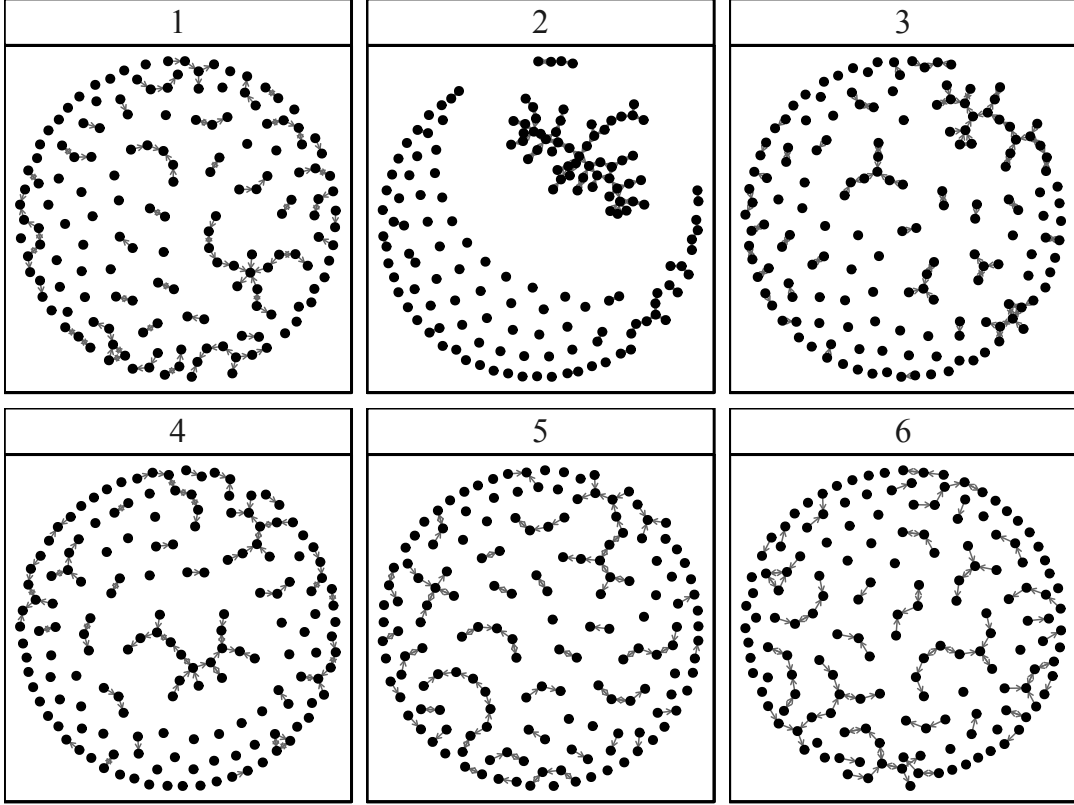
Figure 8: One repitition of a goodness-of-fit lineup testing model M7. The senate data are shown in panel two, and it is evident that none of the other five panels, which show data simulated from model M7, come close to creating the large connected component that is central to the structure of the senate data.

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 results in a more comprehensive goodness-of-fit test for network models.

4.3 Visual Power

The results from the visual power part of our experiment are shown in Figure 9. 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 from 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 $\beta_3, \ldots, \beta_6$: as the parameter value approaches 0, fewer

participants identified the alternative plot. Similarly, as β_1, β_2 approach their estimated values $\hat{\beta}_1, \hat{\beta}_2$, fewer people are able to identify the alternative plot.

We model identification of the alternative data in the lineup by 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 visual power is the probability of detecting the parameter in a lineup of size six. The response variable in our model, $X_{k\ell qr}$, is binary, indicating whether participant r picked the alternative data plot in lineup type ℓ, rep q, for effect k, and it follows a Bernoulli distribution with probability π_{klqr}. We model logit(π_{klqr}) as a sum of the effects for parameter and lineup type, plus random effects for each lineup, δ_{klq}, and for each participant, ϵ_r. 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. The complete hierarchical model is given in Equation 6, where $k \in \{1, 2, 3, 4, 5, 6\}$ corresponds to the effects $\beta_1, \ldots, \beta_6$, respectively, $\ell \in \{-1, 1\}$, and $q \in \{1, 2, 3\}$.

$$X_{klqr} \sim \text{Bernoulli}(\pi_{klqr})$$
$$\text{logit}(\pi_{klqr}) = \eta_{kl} + \gamma_{kl}v + \delta_{klq} + \epsilon_r$$
$$\delta_{klq} \overset{iid}{\sim} N(0, \sigma_\delta^2) \tag{6}$$
$$\epsilon_r \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$$

We fit the hierarchical model with `glmer` from the `lme4` package (Bates et al., 2015). The parameter estimates, standard errors, p-values, and the odds ratio multipliers from the model are summarized in Table **??**. 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

$$E[\text{logit}(\pi_{kl})] = \eta_{kl} + \gamma_{kl}v \tag{7}$$

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

$$\pi_{kl} = \frac{\exp\{\alpha_{kl} + \gamma_{kl}v\}}{1 + \exp\{\alpha_{kl} + \gamma_{kl}x\}} \tag{8}$$
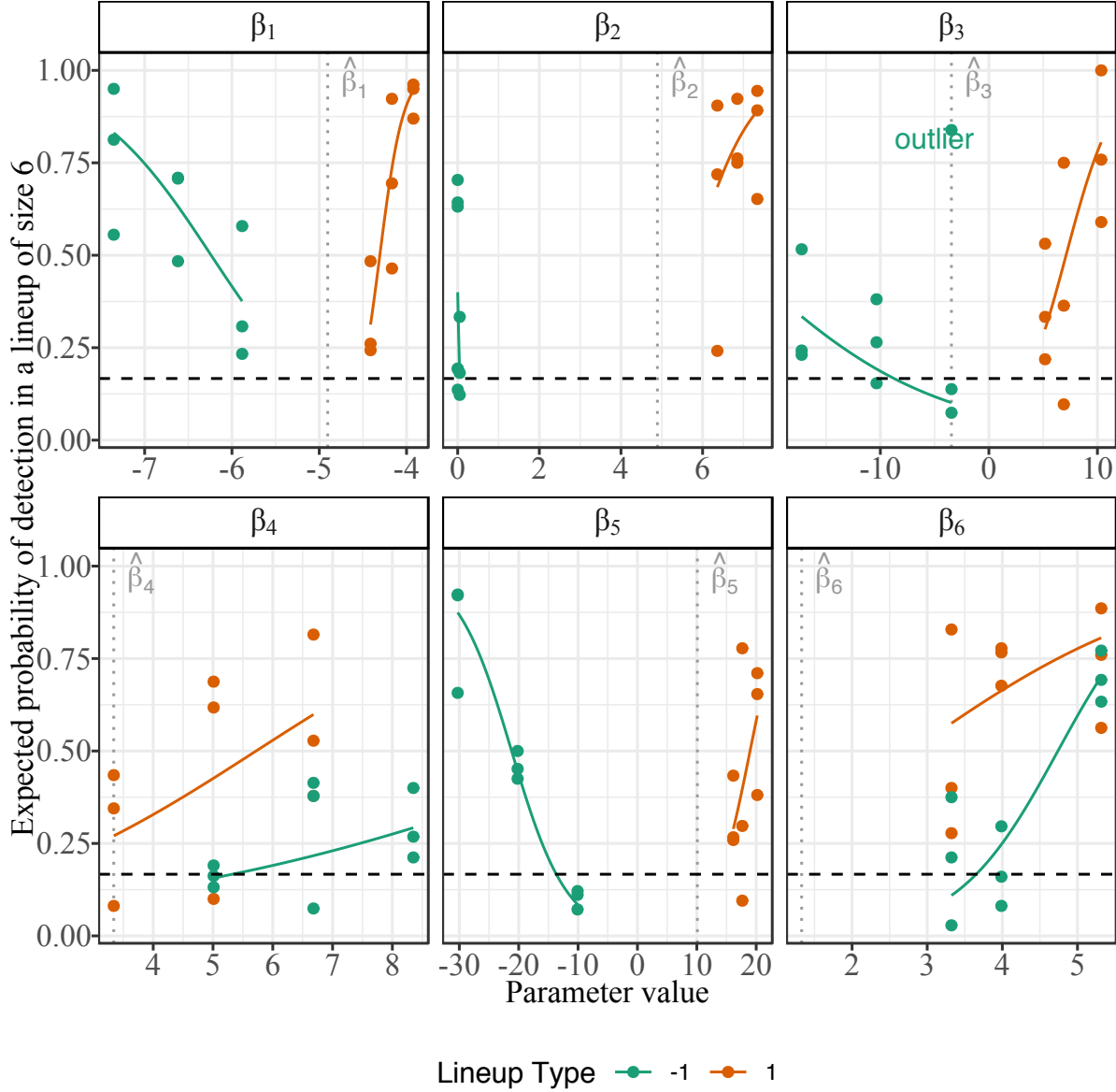
24

Figure 9: Predictions from our GLMM given in Equation 6. 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.

In Figure 9, we see a clear trend in β_2 through β_6 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 2. 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.
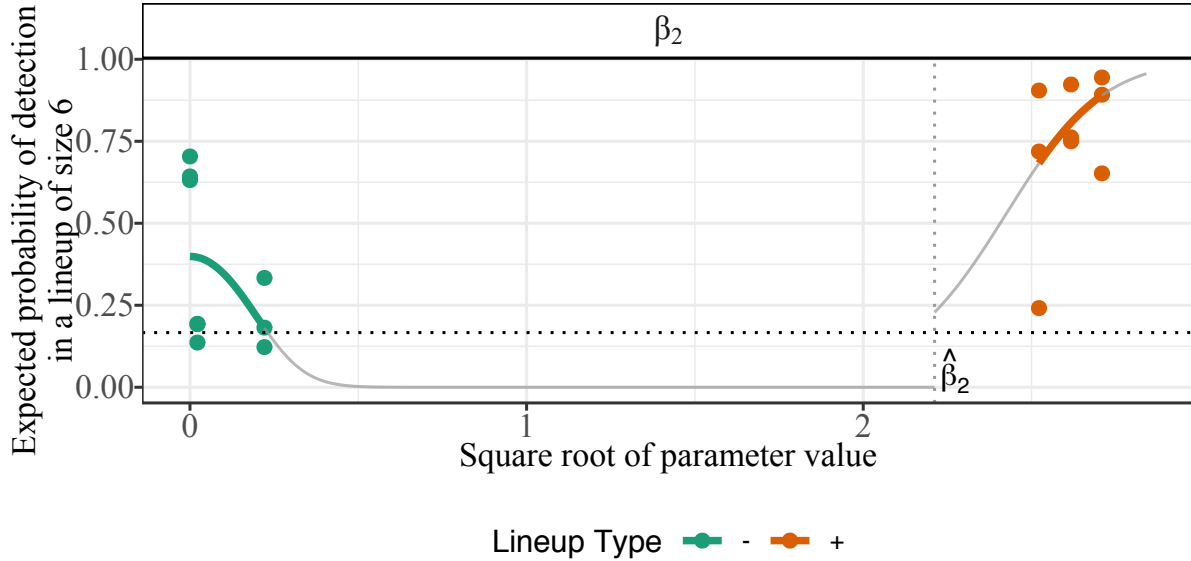

Figure 10: The top middle panel of Figure 9 expanded to show greater detail. The square root of the parameter value is shown on the x-axis. For this parameter, as its value approaches zero, the probability of identifying the alternate data model decreases, then increases, which is noticeably different 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.

We expand portions of Figure 9 in Figures 10-12. These figures show the same prediction regions as in Figure 9, 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 12, 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 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
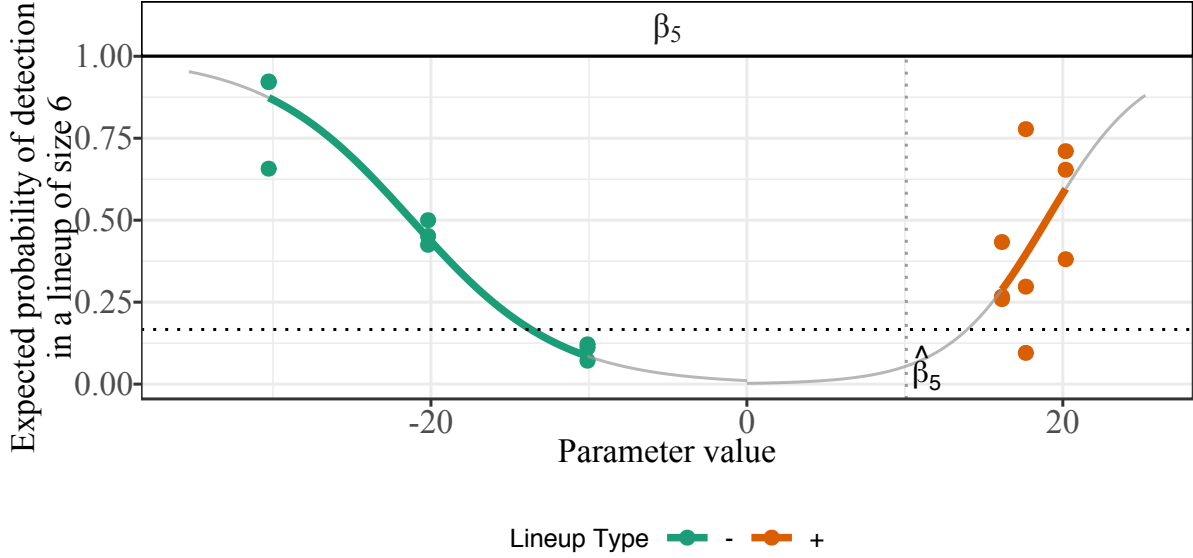
Figure 11: The bottom middle panel of Figure 9 expanded to show greater detail. The parameter value is shown on the x-axis. This parameter most closely follows our hypothesis shown in Figure 2. However, the result is not symmetric. According to the model, people will detect the effect at lower values and with greater frequency as the value increases when it is positive instead of negative.

involves noticing a *lack of structure* as opposed to the presence of *more* structure. We can see a similar effect in Figure 11. 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 lineup types would perform similarly. A lineup type 1 scenario is given in Figure 13, and a corresponding lineup for the lineup type -1 scenario is given in Figure 14, where the parameter of interest is $\beta_4 = 6.681$. 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
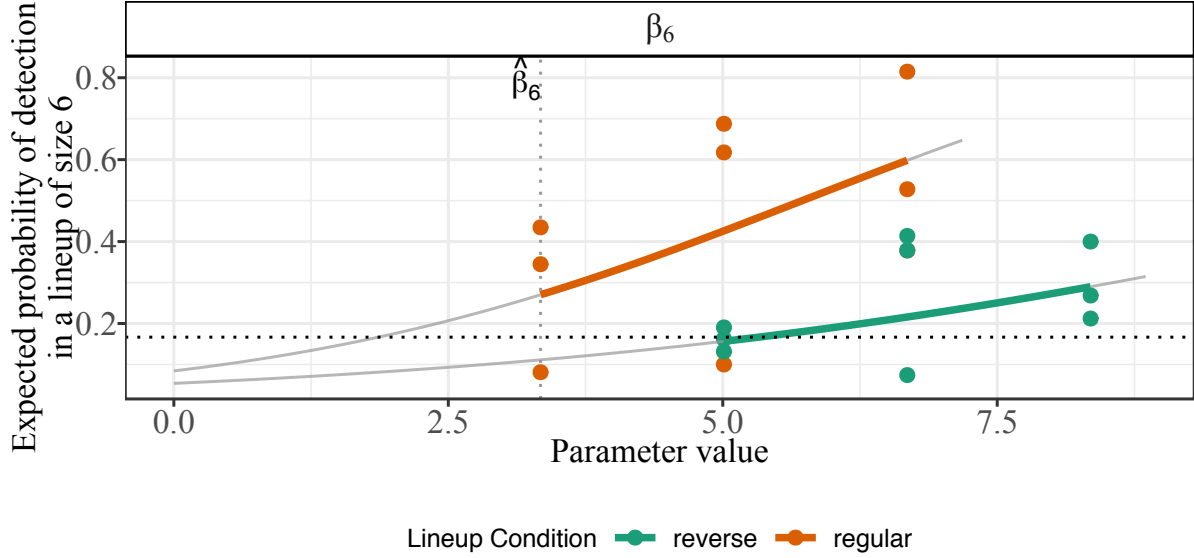
Figure 12: The bottom right panel of Figure 9 expanded to show greater detail. The parameter value is shown on the x-axis. The "reverse" lineup has a much flatter slope than the "regular" lineup, which means the participants had a harder time detecting a more simple M1 structure among many more complex M4 structures. Reversing the lineup scenario was not symmetric as we hypothesized.

five plots from 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. This result is also similar to that of Harrison et al. (2014) because it emphasizes the difficulty of picking out the absence of an effect relative to picking out the presence of an effect. Overall, the visual power of an effect increases as the effect size increases in absolute value, but the power is not symmetric around zero because stong positive effects are noticeable sooner than strong negative effects.

5   Discussion

By using visual inference methods, we have developed new ways to perform significance and goodness-of-fit tests for a complicated family 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
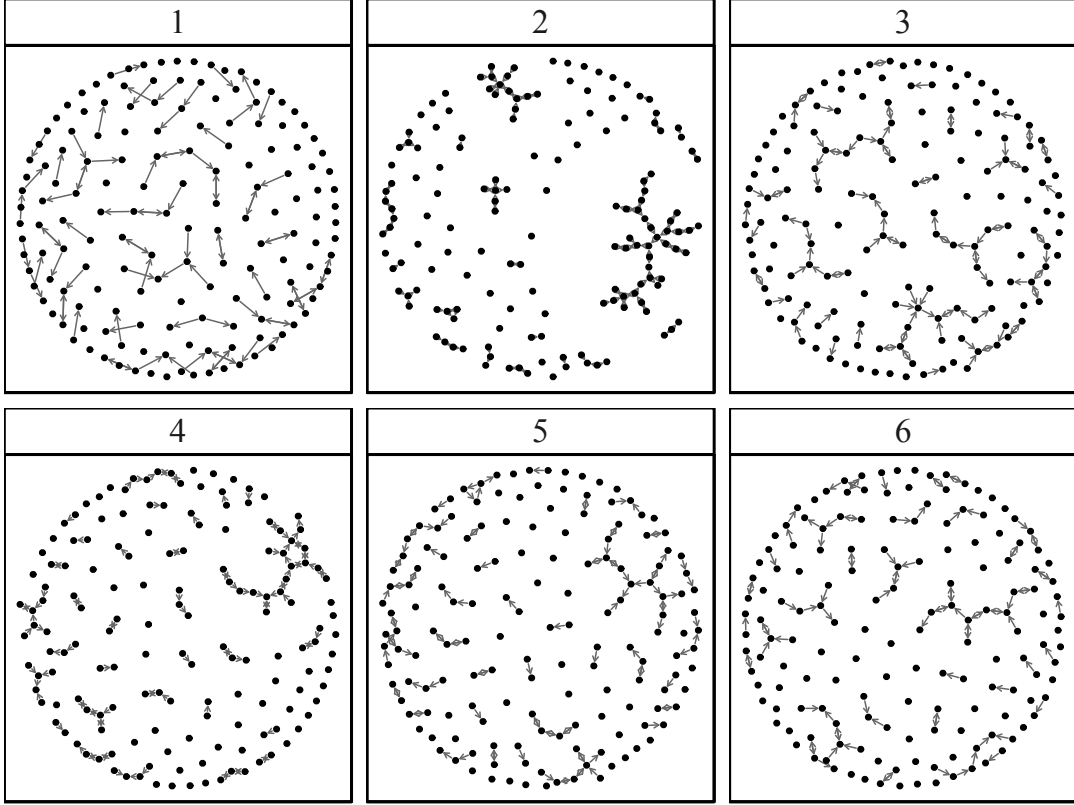
Figure 13: In our experiment, 52.8% of viewers of this plot selected the plot from the alternative model, M4. The "reverse" of this lineup is given in Figure 14, where 41.4% of viewers selected the plot from the alternative model, M1. Here, the alternative plot is $\sqrt{25} - 3$.

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 the model in terms of the *data* itself, not in terms of statistical summaries of the data.

Furthermore, we have found the visual power of some effects in the objective 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
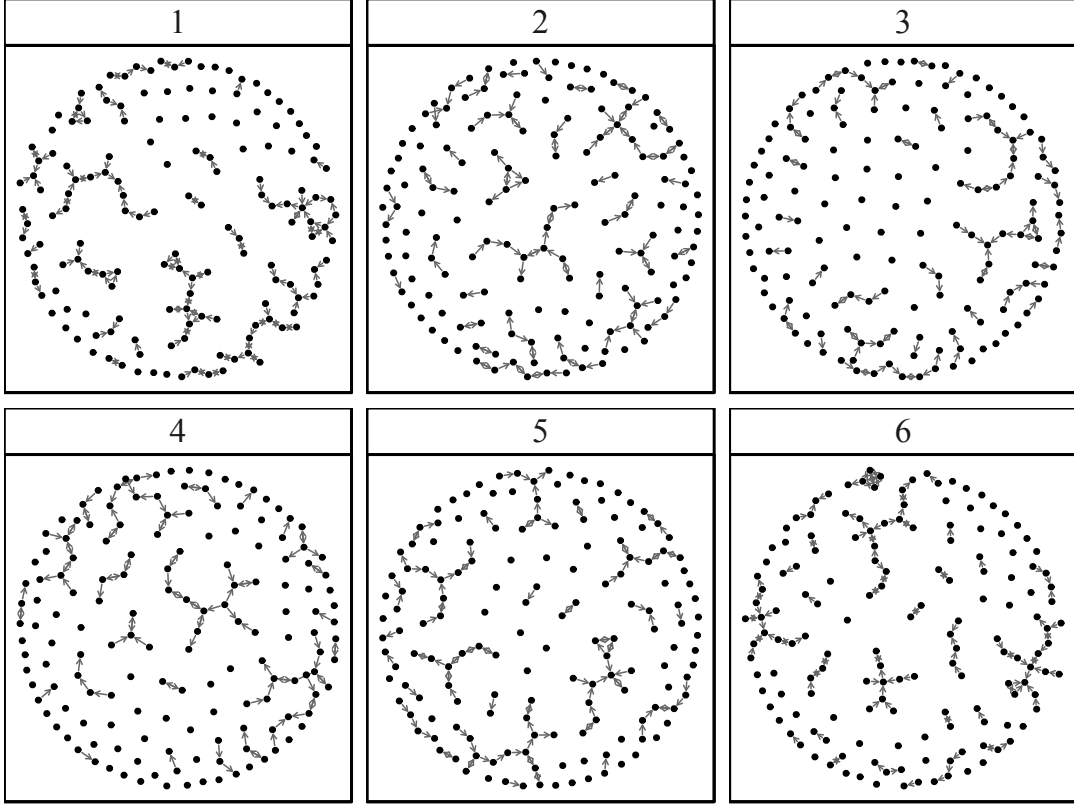
Figure 14: In our experiment, 41.4% people viewing this lineup selected the plot from the alternative model, M1. Here, the alternative plot is $\sqrt{25} - 1$. The other five plots were simulated from model M4.

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 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 hope 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.

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.

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`.

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

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

— (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.