



Bayesian testing of scientific expectations under exponential random graph models

Joris Mulder^{a,*}, Nial Friel^b, Philip Leifeld^c

^a Tilburg University, Netherlands

^b University College Dublin, Ireland

^c University of Essex, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:

Bayesian hypothesis testing

Bayes factors

Exponential random graph models

g-priors

ABSTRACT

The exponential random graph (ERGM) model is a commonly used statistical framework for studying the determinants of tie formations from social network data. To test scientific theories under ERGMs, statistical inferential techniques are generally used based on traditional significance testing using p -values. This methodology has certain limitations, however, such as its inconsistent behavior when the null hypothesis is true, its inability to quantify evidence in favor of a null hypothesis, and its inability to test multiple hypotheses with competing equality and/or order constraints on the parameters of interest in a direct manner. To tackle these shortcomings, this paper presents Bayes factors and posterior probabilities for testing scientific expectations under a Bayesian framework. The methodology is implemented in the R package `BFpack`. The applicability of the methodology is illustrated using empirical collaboration networks and policy networks.

1. Introduction

The exponential random graph model (ERGM) is one of the most widely used statistical models for explaining and predicting the formation of ties between actors in a network using the characteristics of the actors and the endogenous characteristics of the network. ERGMs have resulted in many new insights in various scientific fields, including brain networks in neuroscience (Simpson et al., 2011), board interlocks between firms in management research (Kim et al., 2016), genetic and metabolic networks in biology (Saul and Filkov, 2007), and adolescent friendship networks in sociology (Goodreau et al., 2009).

A statistical analysis under an ERGM is typically executed by first obtaining the maximum likelihood estimates of the network parameters, which quantify the relative importance of the predictor variables, and the corresponding standard errors, which quantify the statistical uncertainty of the estimates given the (limited) information in the network. Subsequently, to test which predictor variables have a statistical effect on the formation of ties in the network, classical p -values are often used. When testing a null hypothesis of whether an ERGM parameter equals zero or not, the p -value is defined as the probability of observing effects that are at least as extreme as the estimated effect from the observed network under the assumption that the true effect equals zero. Despite the ubiquity of this approach in scientific practice however, there is an increasing debate about the usefulness of p -values for statistical hypothesis testing (Benjamin et al., 2018; Mulder and

Wagenmakers, 2016; Rouder et al., 2009). For applied researchers, a potential issue may be its difficult interpretation as p -values are often misinterpreted as error probabilities (Greenland et al., 2016). Besides this practical issue, the p -value also has more fundamental limitations, which also restrict its usefulness for applied social network research under the ERGM framework.

A first limitation is that p -values cannot be used to quantify the evidence in the data in favor of a null hypothesis, but they can only be used to falsify a null hypothesis. This is a limitation when the null hypothesis reflects an important scientific expectation or theory. Leifeld and Schneider (2012), for instance, investigated whether political actors with similar policy preferences form information exchange ties in policy networks. They hypothesized that preference similarity had no additional effect on the formation of ties when covariates based on institutional, relational, and social opportunity structures were also present in the model. When they tested this hypothesis, a p -value larger than the significance level was obtained. The difficulty with this non-significant outcome is that on the one hand the authors could not claim that there is enough evidence in the data to reject the null, and on the other hand it could not be stated that there is evidence in favor of the null (because using a p -value we can only falsify a null). A cause of this problem is that the p -value cannot distinguish between the absence of evidence in the data and evidence of absence of an

* Corresponding author.

E-mail address: j.mulder3@tilburguniversity.edu (J. Mulder).

effect (Dienes, 2014). Therefore, non-significant results cannot guide us about conclusions regarding scientific theories.

A second limitation is that the p -value is statistically inconsistent when the null is true. This is a direct consequence of the key property that the p -value is uniformly distributed under the null. This property is crucial in order to control the type I error probability of incorrectly rejecting the null using a prespecified significance level. A consequence is, however, that there is always a fixed probability (equal to the significance level) of incorrectly rejecting a true null hypothesis, even when the size of the sampled network goes to infinity. Practically, this implies that when collecting information from a huge network (which may be very costly and time consuming) and when there is no effect of a key predictor variable on the formation of ties, there is still a strictly positive chance to draw an incorrect conclusion, which is typically equal to .05.

A final limitation we mention here is that p -values are not designed for testing multiple related hypotheses. Thus, when multiple hypotheses are formulated based on competing scientific expectations, it is not straightforward to test these hypotheses in a direct manner against each other using p -values. Moreover, when a hypothesis also contains order constraints on the parameters of interest, which is often the case in applied research as expectations are often formulated using ‘larger than’ or ‘smaller than’ statements (Hojtink, 2011), p -values can only be used for a limited set of hypothesis tests (Silvapulle and Sen, 2004). For example, p -values are not available for directly testing nonnested order-constrained hypotheses, e.g., $H_1 : \beta_1 \leq \beta_2 \leq \beta_3$ versus $H_2 : \beta_1 > \beta_2 > \beta_3$. Testing order-constrained hypotheses using post-hoc tests is also problematic for another reason: We can obtain conflicting conclusions (e.g., ‘ $\beta_1 = \beta_2$ ’ and ‘ $\beta_1 = \beta_3$ ’ cannot be rejected but ‘ $\beta_2 = \beta_3$ ’ can be rejected), which is problematic for applied statistical practice. Nevertheless, testing order-constrained hypotheses is very relevant for social network research. For example, Leifeld (2018) investigated the effect of collegial relationships on the tendency to co-author papers in scientific collaboration networks. Supervisor–supervisee pairs were expected to show a stronger tendency to collaborate on publications than colleagues only working in the same team. In turn, same-team colleagues were expected to exhibit a stronger tendency to collaborate than colleagues merely working in the same institution. And finally, people in the same institution were expected to co-author at higher rates than people affiliated with different institutions. In such cases, inferential techniques for order-constrained hypotheses are desirable.

To resolve these issues and limitations of p -values, this paper proposes a flexible Bayesian test under the ERGM framework. The methodology uses Bayes factors and posterior probabilities, which (i) can be used to quantify the relative evidence in the data in favor of a null hypothesis, (ii) are consistent under general conditions, and (iii) are broadly applicable for testing multiple hypotheses simultaneously and for testing hypotheses with order constraints on the parameters of interest. A weakly informative ‘unit-information prior’ is proposed which builds on the well-known g -prior approach of Zellner (1986). This prior can be used in an automatic fashion for Bayesian hypothesis testing and estimation without requiring external prior information. Note that prior specification under Bayesian ERGMs is still a largely unexplored topic in the ERGM literature (a notable exception is the conjugate prior of Yin and Butts (2022) for Bayesian estimation). The computation of the resulting Bayes factors is fast due to its reliance on Gaussian approximations of the posterior (following large sample theory). Furthermore, Bayes factors can be computed for any number of hypotheses using only one run of the `bergm` function from the `Bergm` package (Caimo and Friel, 2014). To facilitate its usability, the methodology is implemented in the R package `BFpack` (Mulder et al., 2021b). The main function `BF` only requires a fitted (classical or Bayesian) ERGM using either the `ergm` package (Hunter et al., 2008; Krivitsky et al., 2023) or the `Bergm` package (Caimo and Friel, 2014).

In summary, the current paper makes three important contributions to the literature on Bayesian ERGMs. First, a novel prior specification

method is proposed which can be used in an automatic fashion. Second, a fast method is proposed for computing Bayes factors between many hypotheses. Third, a broad class of multiple hypothesis tests can be executed using equality and/or order constraints on the ERGM coefficients.

The paper is organized as follows. Section 2 provides some background information about ERGMs and about the formulation of hypotheses which may involve order constraints on the parameters of interest. Two motivating empirical applications are then discussed in the context of scientific collaboration networks and information exchange in policy networks. Section 3 describes the methodology for Bayesian hypothesis testing under ERGMs and the implementation in the R package `BFpack` (Mulder et al., 2021b). In Section 4, the numerical behavior of the test is illustrated when testing a null hypothesis (and compared with classical p -values) and when testing order-constrained hypotheses. In Section 5, we apply the methodology in two empirical applications. Section 6 ends the paper with concluding remarks and a discussion.

2. Statistical hypotheses under the ERGM framework

Under an ERGM, the probability of an adjacency matrix \mathbf{Y} of a network of N actors, where $Y_{ij} = 1$ denotes a tie (or edge) between actors i and j , and $Y_{ij} = 0$ denotes the absence of a tie between i and j , is given by

$$p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}) = \frac{\exp\{\boldsymbol{\beta}^\top \mathbf{s}(\mathbf{Y}, \mathbf{X})\}}{\sum_{\mathbf{Y}' \in \mathcal{Y}} \exp\{\boldsymbol{\beta}^\top \mathbf{s}(\mathbf{Y}', \mathbf{X})\}}, \quad (1)$$

where \mathcal{Y} denotes the set of all possible adjacency matrices that can be observed and $\mathbf{s}(\mathbf{Y}, \mathbf{X})$ denotes a vector of K sufficient statistics which are assumed to explain the connectivity in the network given the endogenous characteristics of the network \mathbf{Y} . The sufficient statistics within \mathbf{s} could include the tendency to form ties or the tendency to form triangles as well as the characteristics of the actors summarized in \mathbf{X} . The coefficients in the vector $\boldsymbol{\beta}$ of length K quantify the relative importance of the K sufficient statistics in the formation of ties in the network.

An insightful way to formulate an ERGM is by considering the conditional probabilities of forming a tie for dyad (i, j) given the rest of the graph (Wasserman and Pattison, 1996). The ERGM then comes down to a linear model for the log odds that a tie is present:

$$\log \left(\frac{\Pr(Y_{(ij)} = 1 | \mathbf{Y}_{-(ij)}, \boldsymbol{\beta}, \mathbf{X})}{\Pr(Y_{(ij)} = 0 | \mathbf{Y}_{-(ij)}, \boldsymbol{\beta}, \mathbf{X})} \right) = \log \left(\frac{\Pr(\mathbf{Y}_{-(ij)}^+ | \boldsymbol{\beta}, \mathbf{X})}{\Pr(\mathbf{Y}_{-(ij)}^- | \boldsymbol{\beta}, \mathbf{X})} \right) = \boldsymbol{\beta}^\top \boldsymbol{\delta}_{(ij)}(\mathbf{Y}, \mathbf{X}) \quad (2)$$

where $\mathbf{Y}_{-(ij)}$ denotes the graph \mathbf{Y} where the tie of dyad (i, j) is omitted, and $\mathbf{Y}_{-(ij)}^+$ and $\mathbf{Y}_{-(ij)}^-$ denote the graph where a tie is present and absent for dyad (i, j) , respectively, and $\boldsymbol{\delta}_{(ij)}(\mathbf{Y}, \mathbf{X})$ is the vector of change statistics depending on whether a tie is present or absent for dyad (i, j) , i.e.,

$$\boldsymbol{\delta}_{(ij)}(\mathbf{Y}, \mathbf{X}) = \mathbf{s}(\mathbf{Y}_{-(ij)}^+, \mathbf{X}) - \mathbf{s}(\mathbf{Y}_{-(ij)}^-, \mathbf{X}). \quad (3)$$

In practice, applied network researchers are often interested in testing the relative importance of the sufficient statistics to form ties in the network. Such hypotheses can be translated to competing hypotheses with equality and/or order constraints on the coefficients of the following form:

$$H_t : \mathbf{R}_t^E \boldsymbol{\beta} = \mathbf{0} \ \& \ \mathbf{R}_t^O \boldsymbol{\beta} > \mathbf{0}, \quad (4)$$

for $t = 1, \dots, T$, where the rows of the matrix for the equality constraints, \mathbf{R}_t^E , and rows of the matrix of the one-sided (or order) constraints, \mathbf{R}_t^O , are permutations of $(1, 0, \dots, 0)$, $(-1, 0, \dots, 0)$, or $(1, -1, 0, \dots, 0)$. For example, in the case of a vector of 4 ERGM parameters, i.e., $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \beta_3, \beta_4)$, the matrix multiplication of $(0, 1, 0, 0) \boldsymbol{\beta} > 0$ would correspond to the one-sided constraint $\beta_2 < 0$, and the matrix

multiplication of $(0, -1, 0, 1)\beta > 0$ would correspond to $\beta_4 > \beta_2$. If a hypothesis only contains equality constraints, the matrix \mathbf{R}_1^O is not specified. The same holds if only order constraints are formulated, and then \mathbf{R}_1^E does not need to be formulated. Below, we provide more examples of these matrices for two empirical social network applications.

Application A: Information exchange in policy networks

The first motivating application is the question whether political actors who share similar policy beliefs and political preferences also exchange information in policy networks. Leifeld and Schneider (2012) employed an ERGM to examine the institutional, relational, and social opportunity structures guiding the exchange of information among political actors in a national-level policy network on a contested issue. A policy network is comprised of the interest groups, government agencies, scientific bodies, and other organizations exerting influence over the policy formulation and decision-making process. Actors can exchange technical and scientific or political and strategic information. Explaining under what conditions actors collaborate is instrumental for understanding why certain bills are adopted and why they contain specific elements, for example leaning towards interest group positions.

Leifeld and Schneider (2012) found that any two actors' similarity in policy preferences explained political/strategic information exchange between them. But this effect vanished when opportunity structures like shared committee memberships or shared partners were included in the model. This was seemingly at odds with the prior literature, which had posited that collaboration takes place at higher rates within ideologically homogeneous regions of the network, not across political divides (e.g., Henry et al., 2011). Would the lack of statistical significance once control variables were included have been evidence against an association between preference similarity and tie formation? Leifeld and Schneider (2012) concluded that the effect of preference similarity was "absorbed" by opportunity structures, meaning that opportunity structures were mainly conducive to tie formation between actors with similar preferences, hence preferences per se were no longer a significant predictor themselves. The original study did not have access to a hypothesis testing framework that would have permitted to quantify evidence in favor of the hypothesis that no association existed, controlling for other effects. Below in Section 5, we review this question using the original specification and test the hypothesis that there was no effect of preference similarity on information exchange in the presence of control variables.

H₁: No effect of preference similarity. The first hypothesis assumes that preference similarity has no effect on political/strategic information exchange in a policy network while controlling for the number of edges, governmental target, scientific source, number of common committees, scientific communication, interest group homophily, influence attribution, geometrically weighted edge-wise shared partners (GWESP with a decay parameter of 0.1), geometrically weighted dyad-wise shared partners (GWDSP with a decay parameter of 0.1), and reciprocity. Mathematically, the hypothesis can be written as

$$H_1 : \beta_{\text{pref sim}} = 0$$

Furthermore, the full vector is given by

$$\beta = (\beta_{\text{edges}}, \beta_{\text{pref sim}}, \beta_{\text{gov target}}, \beta_{\text{sci source}}, \beta_{\text{committee}}, \beta_{\text{sci comm}}, \beta_{\text{int group homophily}}, \beta_{\text{infl attr}}, \beta_{\text{GWESP}}, \beta_{\text{GWDSP}}, \beta_{\text{reciprocity}})^T.$$

Consequently, equality-constrained hypothesis $H_1 : \beta_{\text{pref sim}} = 0$ can be written using the matrix notation in (4) as $\mathbf{R}_1^E = (0, 1, 0, \dots, 0)^T$.

H₂: Non-zero effect of preference similarity (complement). The second hypothesis assumes that preference similarity has a non-zero effect on information exchange while controlling for the same covariates as under H_1 . Mathematically, the hypothesis can be written as

$$H_2 : \beta_{\text{pref sim}} \neq 0.$$

Note that even though hypotheses H_1 and H_2 are often denoted by the null hypothesis H_0 and the alternative hypothesis H_1 , respectively, we use the notation in (4) because the methodology is not limited to testing the traditional null and alternative hypothesis.

Based on the analysis of Leifeld and Schneider (2012), a p -value was observed that fell above the significance threshold of .05 resulting in non-significant effect. A non-significant effect, however, does not imply there is evidence that the effect is zero because p -values can only be used for falsifying a null hypothesis. Furthermore, a non-significant finding may be a consequence of an underpowered study. Thus, based on the reported non-significant effect, it is unclear whether there is support for a zero effect of preference similarity or whether the data are inconclusive about which hypothesis is supported by the data.

A secondary question of interest was a specific ordering of effect sizes for different theoretical elements. Leifeld and Schneider (2012) were primarily interested in opportunity structures, such as shared committees, followed by influence attribution, followed by preference similarity. We will explore in Section 4 whether this theoretical importance ranking of effects translates into an order of effect sizes by testing this hypothesis against its equality-constrained counterpart, and against the complement:

$$H_1 : \beta_{\text{committee}} > \beta_{\text{infl att}} > \beta_{\text{pref sim}}$$

$$H_2 : \beta_{\text{committee}} = \beta_{\text{infl att}} = \beta_{\text{pref sim}}$$

$$H_3 : \text{"neither } H_1, \text{ nor } H_2."$$

(5)

Application B: Institutional and functional overlap in co-authorship networks

The second motivating application is the nesting of co-authorship ties in institutional settings. Leifeld (2018) used an ERGM to model the co-authorship networks among all political scientists in Germany and Switzerland during a five-year period. Of particular interest in the analysis was the question whether nomological and idiographic researchers, with their different publication norms, would find themselves separated into distinct communities in the co-authorship network. To answer this question, the study also took into account institutional incentives to collaborate: The more functional overlap existed between any two researchers, the greater the expected odds of collaboration between them, ranging from being in the same discipline and country, working at the same institution, being on the same team, all the way down to direct supervision relationships. But due to the lack of a nested hypothesis testing framework, the original study formulated the effect of increasing levels of institutional and functional overlap as three separate hypotheses, one for each level of functional overlap, without regard for the nested nature of the hypotheses. It concluded from increasingly positive coefficients that more functional overlap must lead to greater odds of collaboration. Here, we evaluate the nested nature of expected collaboration vis-a-vis alternative hypotheses stating that all levels of functional overlap have identical positive effects on co-authorship tie formation and that none of the different functional overlaps exerts any effect on tie formation.

H₁: Increasingly positive effect of functional overlap. The first hypothesis posits that researchers with increasing functional overlap collaborate at higher rates: Researchers with affiliations at the same institution (e.g., university or research institute) are hypothesized to have greater odds of forming collaboration ties than researchers without a shared affiliation. Researchers were expected to form ties at even higher rates within teams, such as chair groups or junior research groups, than

within institutions, compared to pairs of researchers without a shared affiliation. Within teams, researcher pairs where one of the researchers was a professor (usually assistant, associate, or full professor) and the other one was not a professor (usually postdoctoral researcher or Ph.D. student) were expected to have greater odds of co-authorship than the wider team they were embedded in, relative to the reference group of not sharing an affiliation. Mathematically, the hypothesis can be formulated as

$$H_1 : \beta_{\text{supervision}} > \beta_{\text{same team}} > \beta_{\text{same affiliation}} > 0$$

Like in the original study, we control for the number of edges in the network, having shared partners (GWESP with a decay of 0.3), degree distribution (GWDegree with a decay of 0.4), seniority, gender, gender homophily, geographic distance, topic similarity, the share of English articles among an author’s publications, the number of publications, and the similarity between two researchers in their share of English article. Given the notation of a constrained hypothesis in (4), the vector of coefficients is given by

$$\beta = (\beta_{\text{edges}}, \beta_{\text{supervision}}, \beta_{\text{same team}}, \beta_{\text{same affiliation}}, \beta_{\text{GWESP}}, \beta_{\text{GWDegree}}, \beta_{\text{seniority}}, \beta_{\text{gender}}, \beta_{\text{gender homophily}}, \beta_{\text{geographic distance}}, \beta_{\text{topic similarity}}, \beta_{\text{share Eng. art.}}, \beta_{\text{number publications}}, \beta_{\text{Eng. art. similarity}})^T.$$

Thus, hypothesis $H_1 : \beta_{\text{supervision}} > \beta_{\text{same team}} > \beta_{\text{same affiliation}} > 0$ in Application B can be written using the matrix notation in (4) as

$$\mathbf{R}_1^O = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}$$

H₂: Positive and constant effect of functional overlap. The second hypothesis posits that having the same affiliation is the only (positive) factor that plays a role in the tendency to publish together. This hypothesis suggests that affiliations serve as large collaboration communities where there is no additional collaboration effect for researchers who form a supervisor–supervisee pair or for researchers who work on the same team (and who are not in a supervisor–supervisee pair). Mathematically, the hypothesis can be formulated as

$$H_2 : \beta_{\text{supervision}} = \beta_{\text{same team}} = \beta_{\text{same affiliation}} > 0$$

We control for the same covariates as under H_1 . The matrix of equality constraints under this hypothesis is then given by

$$\mathbf{R}_2^E = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots & 0 \end{bmatrix},$$

and the matrix for the one-sided constraint under this hypothesis is equal to

$$\mathbf{R}_2^O = [0 \quad 0 \quad 0 \quad 1 \quad 0 \quad \dots \quad 0].$$

Note that the same hypothesis would be formulated if \mathbf{R}_2^O had a one at the second or third column, and zero elsewhere, because the second, third, and fourth effect are restricted to be equal in \mathbf{R}_2^E .

H₃: No functional overlap effect. The third hypothesis assumes there is no effect of any level of functional overlap, which implies that researchers who form a supervisor–supervisee pair, researchers on the same team (and who do not form a supervisor–supervisee pair), researchers having the same affiliation (and who are not members of the same team), and researchers who do not have the same affiliation all have the same average tendency to publish together. Mathematically, the hypothesis can be formulated as

$$H_3 : \beta_{\text{supervision}} = \beta_{\text{same team}} = \beta_{\text{same affiliation}} = 0$$

We control for the same covariates as under H_1 . The matrix of equality constraints under this hypothesis is then given by

$$\mathbf{R}_3^E = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & \dots & 0 \end{bmatrix}.$$

H₄: Complement. The fourth, complement hypothesis that none of the assumptions under H_1 , H_2 , or H_3 is true. We write this hypothesis as

$$H_4 : \text{neither } H_1, H_2, \text{ nor } H_3.$$

We control for the same covariates as under H_1 . Note that hypothesis H_4 covers the parameter space of the coefficients β that does not satisfy any of the constraints under H_1 , H_2 , and H_3 .

3. Bayesian hypothesis testing

3.1. Background information

The proposed Bayesian hypothesis test is based on the marginal (or integrated) likelihood of the observed network under hypothesis H_t , which is defined by

$$p(\mathbf{Y}|\mathbf{X}, H_t) = \int p_t(\mathbf{Y}|\beta_t, \mathbf{X})p_t(\beta_t)d\beta_t, \tag{6}$$

where β_t denotes the vector of the free parameters under H_t , and $p_t(\beta_t)$ denotes the prior probability distribution of the free parameters under H_t , which reflects the prior expectation and uncertainty about the free parameters before observing the data. Under an ERGM, the marginal likelihood can be computed using the algorithm of Bouranis et al. (2018) for instance. If we consider two competing hypotheses, the ratio of the respective marginal likelihoods is also known as the Bayes factor. For example, the Bayes factor between hypothesis H_1 and H_2 is defined by

$$BF_{12} = \frac{p(\mathbf{Y}|\mathbf{X}, H_1)}{p(\mathbf{Y}|\mathbf{X}, H_2)}, \tag{7}$$

which quantifies how much more plausible the observed network is under H_1 relative to H_2 . From this definition, it automatically follows that the data were more likely to be observed under H_1 than H_2 if $BF_{12} > 1$ and the data were more likely to be observed under H_2 if $BF_{12} < 1$. Because of this interpretation, the Bayes factor can be used as a relative measure of evidence in the data in favor of one hypothesis against another hypothesis. For example, if $BF_{12} = 10$ for the hypotheses in Application A, this would imply that there is 10 times more evidence in the data for hypothesis H_1 than for the alternative H_2 . Furthermore, it can be seen that Bayes factors satisfy a symmetry property, i.e., $BF_{21} = 1/BF_{12}$, implying that if there is 10 times more evidence for H_1 relative to H_2 , then there is 10 times less evidence in the data for H_2 relative to H_1 . Thus, none of the hypotheses has a special role (unlike the null hypothesis in significance testing). Moreover, the Bayes factor follows a transitivity property, e.g., $BF_{31} = BF_{32} \times BF_{21}$, which implies that if H_3 receives 5 times more evidence than H_2 , and H_2 receives 10 times more evidence than H_1 , then H_3 receives 50 times more evidence than H_1 . To facilitate the interpretation, Fig. 1 shows the continuous line of the evidence between two hypotheses as quantified by the Bayes factor, together with a discretized categorization (Raftery, 1995). Note that this categorization does not have to be applied in a strict sense (as the evidence lies on a continuous scale) but it can be useful for researchers who are unfamiliar with Bayes factors. Finally, we note here that Bayes factors are consistent under mild conditions (e.g. O’Hagan, 1995), which implies that the evidence for the true hypothesis goes to infinity as the sample size grows. Thus, $BF_{12} \rightarrow \infty$ as $N \rightarrow \infty$ if H_1 is true, and $BF_{12} \rightarrow 0$ as $N \rightarrow \infty$ if H_2 is true.

Furthermore, in a Bayesian framework, one can specify prior probabilities for the hypotheses, denoted by $P(H_t)$ for $t = 1, \dots, T$, which reflect the plausibility of the hypotheses (such as scientific expectations

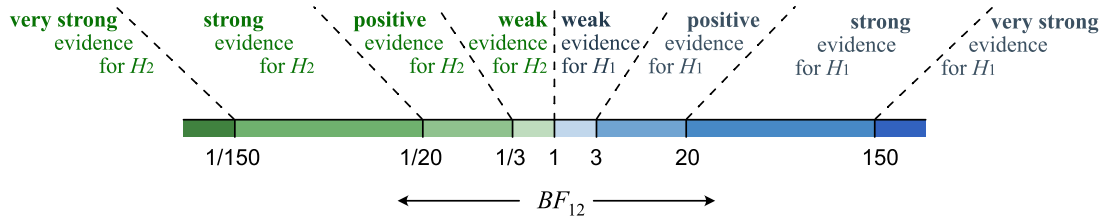


Fig. 1. Guidelines for interpreting the Bayes factor BF_{12} of hypothesis H_1 against H_2 based on (Raftery, 1995).

or substantive theories) before observing the data. A common default choice is the use of equal prior probabilities, i. e., $P(H_t) = T^{-1}$ for $t = 1, \dots, T$. Subsequently, when data are observed, the evidence between the hypotheses in the data is quantified in the Bayes factor, which can be used to update the prior odds of the hypotheses to obtain the posterior odds via

$$\frac{P(H_1|\mathbf{Y}, \mathbf{X})}{P(H_2|\mathbf{Y}, \mathbf{X})} = BF_{12} \times \frac{P(H_1)}{P(H_2)} \quad (8)$$

The posterior probabilities of the hypotheses can then be computed via

$$P(H_t|\mathbf{Y}, \mathbf{X}) = \frac{P(H_t)BF_{t1}}{\sum_{t'=1}^T P(H_{t'})BF_{t'1}}. \quad (9)$$

3.2. Prior specification and Bayes factor computation

In order to compute the marginal likelihoods of the hypotheses and, subsequently, the Bayes factors between the hypotheses, one only needs to formulate the prior distributions for the free parameters under the hypotheses. To simplify this endeavor, we use an encompassing prior approach (Klugkist and Hoijtink, 2007) where an unconstrained (or encompassing) prior, denoted by $p_u(\beta)$, is formulated under the full unconstrained ERGM, denoted by H_u . Truncations are used under the constrained hypotheses according to

$$p_t(\beta) \propto p_u(\beta) \times I(\mathbf{R}_t^E \beta = \mathbf{0} \cap \mathbf{R}_t^O \beta > \mathbf{0}),$$

where $I(\cdot)$ denotes the indicator function. Thus, instead of needing to formulate prior distributions under all T hypotheses, we only need to formulate one unconstrained prior. Another advantage of this choice of the prior is that it substantially simplifies the computation of the Bayes factors as the marginal likelihoods do not have to be computed¹ but instead only the prior and posterior distribution under the full unconstrained ERGM are needed. In fact, the Bayes factor between a constrained hypothesis H_t and the unconstrained model can be written as an extended Savage–Dickey density ratio (e. g. Dickey, 1971; Wetzels et al., 2010; Mulder and Gelissen, 2021):

$$BF_{tu} = \frac{p_u(\beta_t^E = \mathbf{0}|\mathbf{Y}, \mathbf{X})}{p_u(\beta_t^E = \mathbf{0})} \times \frac{P_u(\beta_t^O > \mathbf{0}|\beta_t^E = \mathbf{0}, \mathbf{Y}, \mathbf{X})}{P_u(\beta_t^O > \mathbf{0}|\beta_t^E = \mathbf{0})}, \quad (10)$$

where $\beta_t^E = \mathbf{R}_t^E \beta$ and $\beta_t^O = \mathbf{R}_t^O \beta$. Thus, to compute the first term on the right hand side, we only need to compute the posterior and prior density of the transformed parameter β_t^E at the equality constrained null value. For the second term, we need to compute the posterior and prior probability that the order constraints hold under the condition that the equality constraints hold under the unconstrained model. Note that the posterior density at the null value and the (conditional) posterior probability that order constraints hold serve as measures of relative fit of the constrained hypothesis (relative to the unconstrained model). The prior density at the null value and the prior probability that the order constraints hold serve as measures of relative complexity of

the constrained hypotheses (relative to the unconstrained model). This explicitly shows how the Bayes factor functions as an Occam’s razor by balancing between fit and complexity when quantifying the relative evidence in the data (e.g., Hoijtink et al., 2019).

An unconstrained multivariate normal prior distribution is specified for the ERGM coefficients, which is inspired by the g -prior approach for linear and logistic regression models (Zellner, 1986; Liang et al., 2008; Hanson et al., 2014). The proposed prior can be used in an automatic fashion without requiring manual specification of the prior hyperparameters using external knowledge or simulation techniques. First, we set a noninformative flat prior for the edges parameter as is common for the intercept in the g -prior. This choice ensures that the number of ties in the network is not shrunk towards any (subjectively specified) prior mean. A consequence is that the edges parameter cannot be tested because improper flat (or arbitrarily vague) priors cannot be used for hypothesis testing using Bayes factors (O’Hagan, 1995; Berger and Pericchi, 1996). This is acceptable as the edges parameter typically serves as a nuisance parameter in an ERGM. We can implement it as a normal prior for the edges coefficient using a mean of 0 and a very large variance (e.g., $1e4$). Furthermore, the prior means of the other coefficients are also set to zero. This has two important motivations. First, a prior mean of 0 implies that, a priori, values close to zero are more likely than values far away from zero which implies that smaller effects are on average more likely than (very) large effects (Jeffreys, 1961). Second, as the prior is symmetrical around 0 negative values are equally likely as positive values, which is an objective prior choice for one-sided testing (Mulder et al., 2010). Furthermore, following the g -prior approach, the prior covariance matrix of the ERGM coefficients is based on the implied covariance structure of the predictor variables that is rescaled to one dyadic observation, i. e., $D(\mathbf{\Delta}^T \mathbf{\Delta})^{-1}$, where D is the total number of possible dyads in the network and $\mathbf{\Delta}$ is the design matrix of change statistics (excluding the ones for the edges covariate) under the full ERGM, which is constructed by stacking the row vectors of change statistics $\delta_{(ij)}(\mathbf{Y}, \mathbf{X})^T$ in Eq. (3) over all dyads (i, j) . This results in a weakly informative prior that is dominated by the information in the observed network. Hence, the unconstrained prior can be formulated as

$$p_u(\beta|\mathbf{X}) = N(\mathbf{0}, \text{diag}(1e4, D(\mathbf{\Delta}^T \mathbf{\Delta})^{-1})). \quad (11)$$

Given the multivariate normal prior for the ERGM parameters, a sample from the unconstrained posterior can be obtained in a straightforward manner using the `Bergm` package (Caimo and Friel, 2014). Following large sample theory, a posterior can often be well approximated using a Gaussian approximation (Gelman et al., 2013; Yin and Butts, 2022). Relatedly, note that p -values under ERGMs can also be computed using a Wald-type test by relying on Gaussian approximations. We approximate the unconstrained posterior of the ERGM parameters with a multivariate Gaussian distribution, i. e.,

$$p_u(\beta|\mathbf{X}, \mathbf{Y}) \approx N(\bar{\beta}, \bar{\Sigma}), \quad (12)$$

where $\bar{\beta}$ and $\bar{\Sigma}$ denote the approximated posterior mean and the approximated posterior variance, respectively. Using this approximation, the posterior density and posterior probability in (10) is straightforward, and Bayes factors between the hypotheses of interest can easily be obtained.

¹ Marginal likelihoods can be done using the `evidencePP` or `evidenceCJ` functions in the `Bergm` package (Bouranis et al., 2018; Caimo and Friel, 2013). Priors need to be carefully specified.

3.3. Software implementation in BFpack

The proposed methodology for testing hypotheses using the Bayes factor is implemented in the R package BFpack (Mulder et al., 2021b, version 1.2.3 or later).² This package contains a variety of Bayes factor testing procedures for equality/order-constrained hypotheses for different types of models, including (but not limited to) (multivariate) multiple regression, (multivariate) analysis of (co)variance, generalized linear models, correlational analysis. The main function of the package is BF(). It only requires a fitted model in the form of a R object. Depending on the class of the object, Bayes factors and posterior probabilities are computed under the framework of the fitted model. As part of the current paper, the package is extended to also support objects of class `ergm` or `bergm`. Finally, hypotheses with equality and order constraints on key ERGM coefficients can be formulated using a character string specified in the `hypothesis` argument.

To compute the Bayes factors and posterior probabilities among a set of competing hypotheses under an ERGM, a fitted ERGM object of class `ergm` (from the `ergm` package; Hunter et al., 2008; Krivitsky et al., 2023) or a fitted Bayesian ERGM of class `bergm`³ (from the `Bergm` package; Caimo and Friel, 2014) needs to be placed in the BF function. The same analysis is performed for an object of class `ergm` or `bergm` if the same ERGM is specified using the ‘formula’ argument in these respective functions. The function then extracts the necessary output to compute the Bayes factors and posterior probabilities of interest. Note that when a `bergm` object is plugged in the function (regardless of the prior that was used to obtain this object), a new Bayesian ERGM is fitted using the proposed unit-information prior from which the Bayes factors and posterior probabilities are computed. To get accurate estimates of the Bayes factors and posterior probabilities, it is recommend to collect a sufficient number of posterior draws using the `bergm` function. Throughout this paper we set `main.iter = 100 000` (instead of the default setting of 1000 draws in the `bergm` function, Caimo and Friel, 2014), which can also be included as an argument in the BF function which is then transferred as argument for the `bergm` function. For initial preliminary analyses, the default could be used to avoid long waiting times.

In addition to the confirmatory hypothesis test of multiple constrained hypotheses of the form (4), BFpack also computes the posterior probabilities of a zero effect, a negative effect, or a positive effect by default, i.e.,

$$H_1 : \beta_k = 0 \text{ vs. } H_2 : \beta_k < 0 \text{ vs. } H_3 : \beta_k > 0,$$

for all coefficients β_1, \dots, β_K using equal prior probabilities. The output of these tests can be used for testing the direction of the individual effects on the formation of ties in the network. In Section 5, the software was used to test the empirically informed hypotheses from Section 2.

4. Numerical behavior

Numerical simulations were carried out to study the behavior of the proposed Bayesian test. First, a simulation was executed based on the florentine data (Breiger and Pattison, 1986) to illustrate a key difference between the behavior of the proposed Bayesian test and the classical p -value as a function of the network size. Also, a comparison is made with the BIC and the AIC. Second, simulations were executed to investigate the selection behavior in settings similar to the policy network model of Leifeld and Schneider (2012) with a

² BFpack can be installed from CRAN by running: `install.packages("BFpack")`.

³ The function BF() can only run on a `bergm` object when the network data object is also loaded in the environment because a `bergm` object does not contain the data object, which is required for the analysis.

simplified model specification using predictor variables: edges, preference similarity, influence attribution, shared committees, reciprocity, and geometrically weighted edgewise shared partner. The goals of this second set of simulations were (i) to explore the behavior when testing a null hypothesis and when testing an ordered hypothesis as a function of different true population values and (ii) to illustrate the consistent behavior of the test where the posterior probability of the true hypothesis gradually goes to 1 as the network size grows.

4.1. Illustrating the behavior in comparison to classical p -values

We first illustrate that a Bayes factor behaves differently in comparison to a p -value as a function of the sample (or network) size. The difference in behavior is a consequence of the fact that the Bayes factor is statistically consistent if the simpler (null) model is true while the p -value is not. As the sample (or network) size grows to infinity, the evidence for a true null model as quantified by the Bayes factor grows to infinity (and correspondingly, the posterior probability of the true null goes to 1). A significance test based the p -value on the other hand is statistically inconsistent. There always is a probability (typically .05) to incorrectly reject the null, even when the network size grows to infinity. To illustrate this different behavior, we consider a relatively small network that results in a 2-sided p -value of .05 for an effect of a predictor variable. Next we consider larger networks in such a way that the p -value for this effect remains constant at .05. This is achieved by letting the true parameter value gradually decrease to zero, and thus the null is true in the limit. Consequently, the evidence for the null using the Bayes factor then gradually increases, eventually towards infinity. In the literature, this conflicting behavior is also referred to as the Jeffreys–Lindley paradox (Jeffreys, 1961; Lindley, 1957).

We consider the following simple ERGM based on the Florentine data (Breiger and Pattison, 1986):

$$\text{floomarriage} \sim \text{edges} + \text{kstar}(2) + \text{absdiff}(\text{"wealth"})$$

and the hypothesis test of whether the effect of the absolute difference in wealth, `absdiff("wealth")`, is zero or not, i.e.,

$$H_1 : \beta_{\text{absdiff wealth}} = 0$$

$$H_2 : \beta_{\text{absdiff wealth}} \neq 0.$$

Simulated networks are considered consisting of $n = 7, 10, 15, 20, \dots, 50$, and 55 actors. The parameter values of the edges and `kstar(2)` statistic were set equal to the MLEs of the empirical florentine data (although the exact choices did not qualitatively affect the results), and the effect of `absdiff("wealth")` was set such that the p -values of the generated data was close to .05. The covariate values of the actors were sampled from the empirical wealth values with replacement. Of the generated networks, only the ones that resulted in a two-sided p -value in the interval (0.045,0.055) were stored, resulting in 300 networks for each network size n .

Fig. 2 shows the median MLE (black circle) with confidence bounds and the Bayesian posterior modes (red crosses) with credibility interval bounds (upper left panel), the median of the p -values (upper right panel), which is .05 by construction, the median of the Bayes factors of the simple hypothesis H_1 against the alternative H_2 (middle left panel), the median of the posterior probability of H_1 (assuming equal prior probabilities for H_1 and H_2) (middle right panel), the median of the approximated Bayes factor based on the BIC (lower left panel), and the median of the evidence ratio of H_1 against H_2 based on the AIC⁴ (lower right panel), all as a function of the network size. As the effect gradually decreases towards zero, we see that the evidence for the simpler hypothesis, which assumes that the effect is zero (middle

⁴ The evidence based on the BIC and AIC were computed according to $\exp(-(\text{BIC}(H_1) - \text{BIC}(H_2))/2)$ and $\exp(-(\text{AIC}(H_1) - \text{AIC}(H_2))/2)$, respectively.

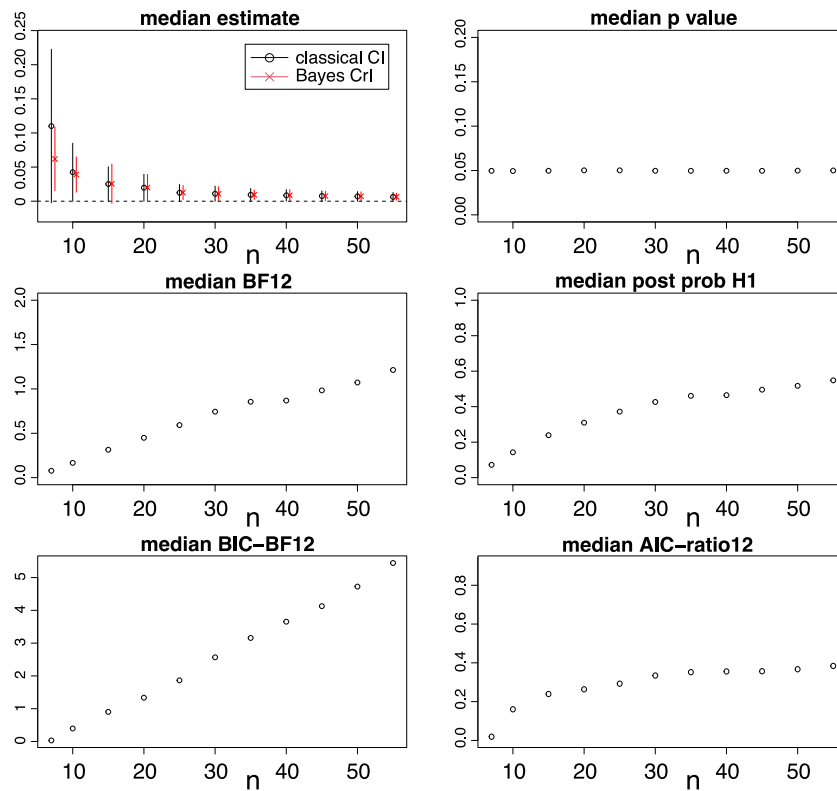


Fig. 2. The median of the estimated effect MLE and posterior mean with confidence (CI) and credible (CrI) bounds, respectively, the median of the p -value, the median of the Bayes factor of $H_1 : \beta_{\text{absdiff}}(\text{wealth}) = 0$ against $H_2 : \beta_{\text{absdiff}}(\text{wealth}) \neq 0$, the median of the posterior probability of H_1 , the median of the evidence based on the BIC, and the median of the evidence based on the AIC based on 300 generated data sets as a function of the network size $n = 7, 10, 15, 20, \dots$, and 55.

left panel), gradually increases. Eventually it results in evidence in favor of the simpler hypothesis H_1 (around $n = 50$) while the p -value remains constant at .05 (by construction). The posterior probability of H_1 shows similar behavior as the Bayes factor as it behaves as a normalized version of the evidence (middle right panel). The behavior of the Bayes factor is a consequence of its consistency, which implies that the evidence goes to infinity towards the true hypothesis as the network size grows. The p -value, on the other hand, is inconsistent (by definition) because it is uniformly distributed under the null regardless of the sample size. For this reason, the p -value also cannot be used as a quantification of the evidence in favor of the null (see also Rouder et al., 2009). The Bayes factor does not have this property as all hypotheses are compared on equal footing. If the evidence clearly points towards H_1 (or H_2), this implies that H_1 (or H_2) was most likely to have generated the data, and if the Bayes factor does not clearly point towards any of the hypotheses (i.e., the Bayes factor is close to 1), then there is absence of evidence and more data are required to obtain clearer evidence towards the hypothesis that is most likely true. This interpretation holds regardless of the network size.

As a comparison, we also provide the evidence ratios based on the BIC (Fig. 2, lower left panel) and the AIC (Fig. 2, lower right panel). Both show increasing trends as a function of the network size, similar to the proposed Bayes factor. However the evidence for the null based on the BIC increases considerably faster when an evidence ratio of approximately 1 is already obtained around $n = 15$. Thus the BIC is considerably more conservative towards the null than the proposed Bayes factor. Note that the BIC is also sometimes regarded as being too conservative towards simpler models (e.g. Weakliem, 1999). To understand the difference in behavior, note that the proposed Bayes factor uses a unit information prior that is centered at 0 while the BIC behaves as a marginal likelihood using a unit information prior that is centered at the MLE (Kass and Wasserman, 1995; Raftery, 1995). Another difference is that the BIC is based on a Laplace approximation of the likelihood while the proposed Bayes factor is based on a

Laplace approximation of a (kernel of the) posterior, which combines information in the unit information prior and the likelihood. We refer the interested reader to Raftery (1995), Weakliem (1999), and Raftery (1999) for relevant discussions about the BIC approximation for hypothesis testing and model selection. Finally, we note that the evidence based on the AIC increases with a slower rate than the proposed Bayes factor, making the AIC more liberal towards larger models. This is related to the property of the evidence ratio based on the AIC which converges to a constant and not infinity (O’Hagan, 1995), which implies inconsistent behavior in the long run as the sample size grows.

4.2. Illustrating the behavior using simulated policy networks

In this section, two hypothesis tests are discussed based on the empirical networks from Section 2. First, a traditional null hypothesis test is considered of whether the coefficient of preference similarity equals zero or not, i. e.,

$$H_1 : \beta_{\text{pref sim}} = 0$$

$$H_2 : \beta_{\text{pref sim}} \neq 0.$$

Second, a multiple order-constrained hypothesis test is considered of a specific ordering of three coefficients versus an equality-constrained alternative versus the complement, i. e.,

$$H_1 : \beta_{\text{pref sim}} < \beta_{\text{infl attr}} < \beta_{\text{committee}}$$

$$H_2 : \beta_{\text{pref sim}} = \beta_{\text{infl attr}} = \beta_{\text{committee}}$$

$$H_3 : \text{neither } H_1, \text{ nor } H_2.$$

Networks were considered of 10 actors, 30 actors (as in the empirical data), and 90 actors. For the network of 30 actors, the population values were set equal to the MLEs of the fitted model to the empirical data. For the network of 10 actors, the population values were set equal to the MLEs based on a subset of the first 10 actors of the empirical

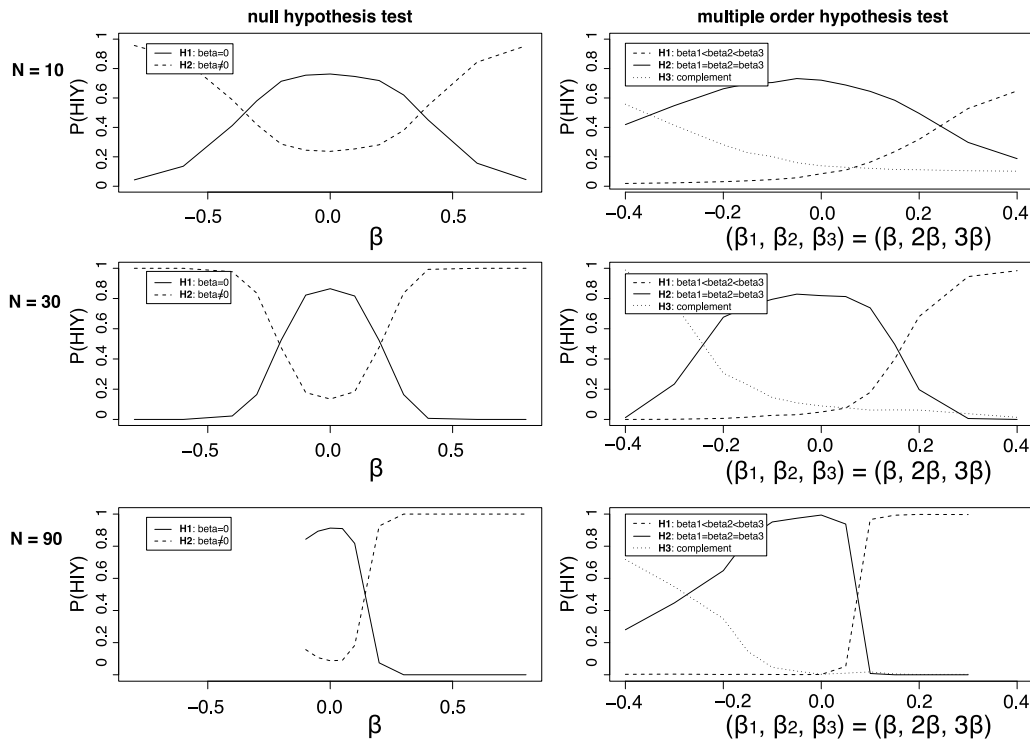


Fig. 3. Median posterior probability of the first precise test (left panels; $H_1 : \beta = 0$ vs $H_2 : \beta \neq 0$) and the multiple order test (right panels; $H_1 : \beta_1 < \beta_2 < \beta_3$ vs $H_2 : \beta_1 = \beta_2 = \beta_3$ vs $H_3 : \text{complement}$) for networks of 10 actors (upper panels), 30 actors (middle panels), and 90 actors (lower panels) based on 100 randomly generated networks for every grid value. For the generated networks with 90 actors, the ERGM could often not be fitted for negative grid values for both tests. For the first test the lines are therefore omitted.

data. For the network of 90 actors, the population values were equal to the MLEs based on the 90×90 block diagonal adjacency matrix where the 30×30 empirical adjacency matrix was placed 3 times on the block diagonal. This setting ensured that the mean degree of the simulated networks was comparable with the empirical network.

To investigate the behavior of the proposed test, only the values of the key parameters that were tested were changed. In the simulation for the first test, the value for $\beta_{\text{pref sim}}$ was varied on a grid from -0.8 to 0.8 . For the second test, network data were generated using values for $(\beta_{\text{pref sim}}, \beta_{\text{infl attr}}, \beta_{\text{committee}})$ equal to $(\beta, 2\beta, 3\beta)$, where β was varied on a grid of -0.4 to 0.4 . Thus, when β is positive, the order-constrained hypothesis H_1 is true, when β is zero the equality hypothesis H_2 is true, and when β is negative, the complement hypothesis H_3 is true. For each grid value, 100 networks were generated, and Bayes factors were computed between the hypotheses and transformed to posterior probabilities. We checked whether the posteriors can be accurately approximated with Gaussian distributions. This was the case over the entire range of values for the tested parameters. The plots looked similar as in Figs. 6 and 7 (discussed later for the empirical applications), and therefore the similarly looking plots for the simulated data are omitted here.

Fig. 3 shows the median posterior probability of the first test (left panels) and the multiple order test (right panels) for networks of 10 actors (lower panels), 30 actors (middle panels), and 90 actors (lower panels). Before discussing the results, we note that for the larger networks with 90 actors, the MLE often did not exist in the lower range of the parameter space in both tests. This seemed to be caused by the generated networks being too sparse. For the first test, the results are therefore omitted for parameter values smaller than -0.2 , and in the second test, only the generated networks for which the MLE existed were used, resulting in more moderate estimates (and, thus, more moderate outcomes of the evidence).

For both tests, the figure shows the anticipated behavior, where on average the true hypothesis receives the largest posterior probability and this increases for larger networks. Furthermore, we see that the

posterior probability for the equality hypothesis (H_1 in the first test and H_2 in the second test) goes to 1 with a slower rate when it is true than the rate when the alternative hypotheses are true. This implies that more data are generally needed to get a large posterior probability of a true equality hypothesis than for unconstrained or order-constrained hypotheses. Intuitively, this seems reasonable because it is easier to find evidence that a parameter is not exactly 0 (or that several parameters are not exactly equal) than to find evidence that a parameter is *exactly* 0 (or that several parameters are *exactly* equal). This behavior is also often observed for Bayes factors for other testing problems.

For the order-constrained hypothesis test, it is interesting to see that smaller effects are needed in the direction of the order-constrained hypothesis to acquire the same posterior probabilities in comparison to the effect in the opposite direction for the complement hypothesis. For example, in the case of a network with $N = 30$ actors (right middle panel of Fig. 3), the posterior probability for the order-constrained hypothesis H_1 is largest when the true value is about 0.16 while the posterior probability for the complement hypothesis H_3 is largest when the effect β is lower than -0.24 . This can be explained by the fact that the order-constrained hypothesis covers a smaller parameter space than the complement hypothesis, and therefore the complement receives a larger penalty for model complexity. In fact, the complexity of order-constrained hypotheses is implemented via the prior probability that the constraints hold (e.g., see Mulder et al., 2010). This illustrates that the criterion also behaves as an Occam’s razor when assessing order-constrained hypotheses.

To see this more clearly, Fig. 4 displays the Bayes factors of order-constrained hypothesis H_1 versus the unconstrained model, which is computed as the posterior probability that the order constraints hold divided by the prior probability that the order constraints hold under the unconstrained model; see Eq. (10). Thus, in the case of overwhelming evidence for the order constraints, the posterior probability that the order constraints hold is approximately 1, and the Bayes factor of the order-constrained hypothesis against the unconstrained model is approximately equal to the reciprocal of the prior probability that the

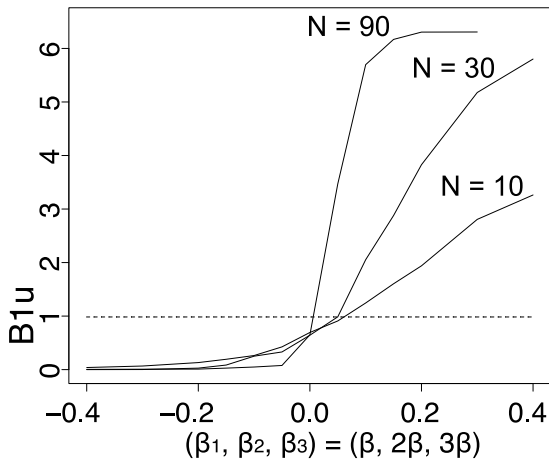


Fig. 4. The median Bayes factor of the order-constrained hypothesis $H_1 : \beta_{\text{pref sim}} < \beta_{\text{infl attr}} < \beta_{\text{committee}}$ against the unconstrained model, which is computed as the ratio of posterior and prior probabilities that the constraints hold, while varying the ordered effect size on the x axis, for different network sizes of $N = 10, 30,$ and 90 actors.

constraints hold. Note that the prior probability depends on the prior covariance structure, which depends on the design matrix via (11). For example, if the covariates are independent, the prior probability that the order constraints $\beta_{\text{pref sim}} < \beta_{\text{infl attr}} < \beta_{\text{committee}}$ hold under the unconstrained model is equal to $1/6$, which can be explained by the fact that there are 6 possible orderings of 3 parameters. In this case, the Bayes factor of the order-constrained hypothesis against the unconstrained model grows to 6 if the ordered effect increases. This is also what can be observed in Fig. 4 for a slightly different prior covariance matrix (based on the data from Leifeld and Schneider, 2012).

5. Empirical applications (revisited)

Application A: Information exchange in policy networks

The main interest was whether preference similarity plays a role in information sharing in policy networks or whether the effect is absorbed by institutional, relational, and social opportunity structures, i.e.,

$$H_1 : \beta_{\text{pref sim}} = 0$$

$$H_2 : \beta_{\text{pref sim}} \neq 0.$$

These hypotheses are tested under Model 2 of Leifeld and Schneider (2012).⁵ The Bayes factor between H_1 and H_2 was computed using Eq. (10) (only using the first part on the right hand side) where the posterior and prior density at 0 are approximately equal to 2.04 and 0.387, respectively, resulting in $BF_{12} = \frac{2.04}{0.387} = 5.2$ (see Fig. 5), which implies that the network was 5.20 times more likely to be generated under a model that assumes that preference similarity equals zero than by a model that assumes that preference similarity can attain any real value. This suggests positive evidence in favor of no effect of preference similarity on tie formation. Note that the computation of the evidence as the posterior and prior ratio at the null value shows an intuitive correspondence with Bayesian estimation: When the posterior density at the null value is high (low), the null value is likely (unlikely) according to the posterior and the evidence in favor of the null hypothesis value is also high (low) (see also Wagenmakers et al., 2010).

When assuming equal prior probabilities for the two hypotheses, i.e., $P(H_1) = P(H_2) = .5$, the posterior probabilities are then given

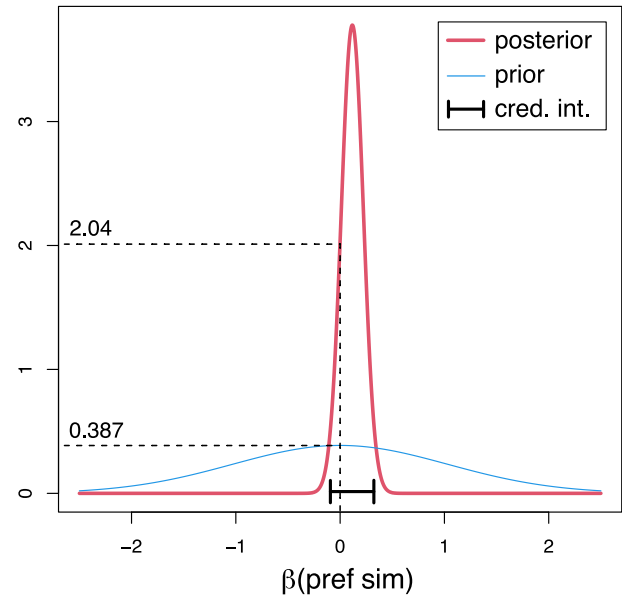


Fig. 5. Approximated posterior (thick red line) and prior (thin blue line) for the effect of preference similarity based on the policy network. The posterior and prior density at the null value are equal to 2.04 and 0.387, respectively, resulting in a Bayes factor of $\frac{2.04}{0.387} \approx 5.2$ in favor of a zero effect. The thick black line denotes the 95% Bayesian credibility interval.

by $P(H_1|Y) = .839$ and $P(H_2|Y) = .161$. Thus after observing the data, there is a posterior probability of approximately 84% for the model that assumes that preference similarity plays no role in tie formation in this policy network. Note that no formal decision needs to be drawn based on these posterior probabilities. But if one would conclude that H_1 is the true hypothesis, there would be a conditional error probability of about 16% of drawing the wrong conclusion given the observed network conditional on the observed network.

Further note again that the two-sided p -value of .136 in comparison to the traditional threshold value of .05 would imply that there was not enough evidence to reject the null. Thus, based on the p -value, we cannot tell whether there is evidence of no effect of preference similarity (because p -values cannot be used as a measure of support for a zero effect) or absence of evidence (i.e., the information in the data is inconclusive). The Bayesian test, on the other hand, removes this uncertainty as its outcome suggests that there is positive evidence that the effect is zero, and thus, that the non-significant result is not a consequence of an underpowered study. The weight of evidence between H_1 and H_2 was also computed using the BIC and the AIC, which resulted in 13.63 (i.e., positive evidence for H_1) and 1.26 (i.e., practically equal evidence between the hypotheses with a very slight preference towards H_1), respectively. These results are in agreement with the behavior we observed in the simulation results presented in Section 4.1, which also showed that the BIC is generally more conservative towards the simpler model and the AIC is more liberal towards the larger alternative.

Besides testing the effect of preference similarity, we also discuss the results for the other coefficients. Table 1 presents the MLEs, the standard errors, the p -values, the Bayesian posterior means, the posterior standard deviations (both based on the proposed unit information prior), and the posterior probabilities when testing whether a coefficient is zero, negative, or positive (assuming equal prior probabilities of $\frac{1}{3}$). The latter posterior probabilities are standard output of BFpack. Note that the Bayesian testing results are not provided for the (nuisance) edges coefficient because an improper flat prior is used for this parameter. When looking at the p -values and the posterior probabilities of the null hypotheses of the other coefficients in Table 2, we see that generally both statistics point in the same direction. However the

⁵ The R code for analyses can be found in Appendix.

Table 1

Classical results (MLE, standard error, and p -value), Bayesian estimates (using the normal unit information prior), and posterior probabilities of a null, a negative, and a positive effect (assuming equal prior probabilities) for all ERGM coefficients of the policy network (Application A).

	Classical results			Bayesian estimates		Posterior probabilities		
	MLE	s.e.	p	post. mean	post. sd	$P(\beta = 0 Y)$	$P(\beta < 0 Y)$	$P(\beta > 0 Y)$
edges	-4.039	1.290	0.002	-2.525	0.675	–	–	–
pref. sim. (st.)	0.118	0.079	0.136	0.116	0.106	0.722	0.037	0.241
reciprocity	0.808	0.248	0.001	0.765	0.298	0.138	0.005	0.857
int. group homophily	1.067	0.291	0.000	1.222	0.485	0.178	0.005	0.817
gov. target	0.597	0.189	0.002	0.561	0.248	0.259	0.009	0.732
sci. source	0.072	0.218	0.742	0.063	0.269	0.822	0.072	0.106
common committees (st.)	0.727	0.122	0.000	0.839	0.173	0.000	0.000	1.000
sci. communication	2.910	0.630	0.000	2.935	1.006	0.039	0.002	0.958
infl. attr. (st.)	0.439	0.088	0.000	0.438	0.114	0.003	0.000	0.997
GWESP(0.1)	2.552	1.129	0.024	1.363	0.597	0.094	0.012	0.895
GWDSP(0.1)	-0.134	0.049	0.007	-0.208	0.087	0.146	0.847	0.007

Table 2

Bayes factors between the hypotheses in the second test in Application A, and posterior probabilities using equal prior probabilities.

Hypotheses	Bayes factors			Posterior probabilities $P(H_i Y)$
	H_1	H_2	H_3	
$H_1 : \beta_{\text{committee}} > \beta_{\text{infl. attr.}} > \beta_{\text{pref. sim.}}$	1.000	74.335	97.953	.977
$H_2 : \beta_{\text{committee}} = \beta_{\text{infl. attr.}} = \beta_{\text{pref. sim.}}$	0.013	1.000	1.318	.013
$H_3 : \text{complement}$	0.010	0.759	1.000	.010

interpretation and theoretical properties of these methods (e.g., see the simulations in Section 4.1) are fundamentally different. This needs to be taken into account when interpreting these results. Furthermore, we can see that most (but not all) posterior means are closer to zero than the MLEs. This could be attributed to some prior shrinkage towards the prior mean of 0.

The secondary question in this application was the test of the order-constrained hypothesis of shared committees followed by influence attribution, followed by preference similarity versus a (null) hypothesis of equal effects versus the complement that neither of these two hypotheses hold; see Eq. (5). The edge covariates were standardized to be able to compare the effects on the same unified scale. The Bayes factors between the hypotheses and the posterior probabilities (when assuming equal prior probabilities of $\frac{1}{3}$ for each hypothesis are given in Table 2). We see that the order-constrained hypothesis receives approximately 74 and 98 times more evidence than the equality hypothesis H_2 and the complement H_3 , respectively. These results also display the largest posterior probability for the order-constrained hypothesis H_1 with approximately 98%. Thus, after observing the data, there is a strong belief that the (standardized) effect of shared committees is strongest, followed by influence attribution, followed by preference similarity. These results are also in agreement with the estimates of the coefficients in Table 1.

We end this application by showing the accuracy of the Gaussian approximation of the posterior of preference similarity, common committees, and influence attribution based on the posterior draws from the bergm function. Fig. 6 plots the estimated posterior (dashed black line) and the Gaussian approximation (red solid line). The figure shows that the posteriors can be well approximated using Gaussian distributions.

Application B: Institutional and functional overlap in co-authorship networks

The interest is in testing competing hypotheses about an increasing positive effect of functional overlap of the form

- $H_1 : \beta_{\text{supervision}} > \beta_{\text{same team}} > \beta_{\text{same affiliation}} > 0$
- $H_2 : \beta_{\text{supervision}} = \beta_{\text{same team}} = \beta_{\text{same affiliation}} > 0$
- $H_3 : \beta_{\text{supervision}} = \beta_{\text{same team}} = \beta_{\text{same affiliation}} = 0$
- $H_4 : \text{complement,}$

which are tested under the Swiss collaboration network data from Leifeld (2018).⁶ The Bayes factors between the hypotheses can be found in the evidence matrix in Table 3. We see that hypothesis H_1 , which assumes an increasing positive collegial effect, receives decisive evidence against H_2 (constant positive collegial effect) and H_3 (no collegial effect), and strong evidence against the complement hypothesis. The resulting posterior probabilities are printed in the last column of Table 3 when using equal prior probabilities. We see that the posterior probabilities for H_1 , H_2 , H_3 , and H_4 are approximately 91%, 2%, 0%, and 7%. Thus, H_1 , which assumes an increasing positive effect of functional overlap, is clearly most likely after observing the data. Furthermore, we can safely rule out hypothesis H_3 , which assumed no functional overlap effect, given the zero posterior probability. Again, we note that no decision needs to be made which hypothesis is true because these outcomes can be interpreted on a continuous scale. But if it was concluded that hypothesis H_1 is true, there would be a conditional probability of drawing the wrong conclusion of about 9% given the observed network.

A visual check was performed to see how well the posterior distribution of the key parameters in this hypothesis test can be approximated with a Gaussian distribution. In Fig. 7, the black dashed lines are the estimated posteriors, and the solid red lines are the Gaussian approximations. Overall, the Gaussian approximations seem accurate with a slight deviation for the posterior of the supervision effect due to its skewness. Based on these results, we conclude that the Gaussian approximation is sufficiently accurate.

Finally, we present the classical results, Bayesian estimates, and posterior probabilities of the hypotheses for all separate coefficients in Table 4. We see some differences between the Bayesian and the classical results. Several coefficients result in relatively small p -values while the Bayesian posterior probability of a zero effect for these coefficients are substantial, such as GWDegree, gender, share English articles, number of publications, and English article similarity. These differences can be explained by the relatively large posterior standard deviations in comparison to the classical standard errors implying more posterior uncertainty regarding the true parameter value. Again, we see that some Bayesian estimates are considerably larger than the classical estimates. The table also shows that the Bayesian estimate of the effect of same team is slightly larger than the estimated effect of supervisor

⁶ The R code for analyses can be found in Appendix B.

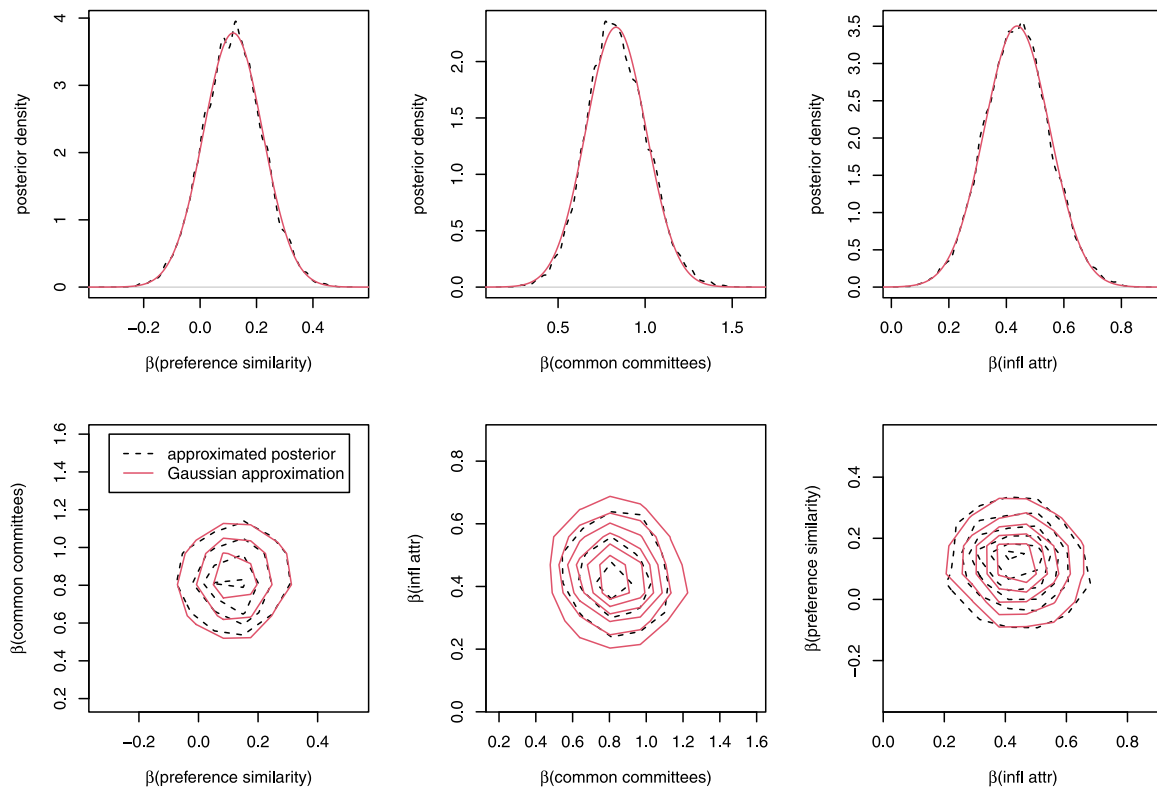


Fig. 6. Posterior density estimate based on 220,000 posterior draws (dashed black line) and the Gaussian approximation of the posterior (solid red line) for the policy network.

Table 3

Bayes factors between the hypotheses in Application B, and posterior probabilities based on equal prior probabilities.

Hypotheses	Bayes factors				Posterior probabilities $P(H_i Y)$
	H_1	H_2	H_3	H_4	
$H_1 : \beta_{supervision} > \beta_{same\ team} > \beta_{same\ affiliation} > 0$	1.000	59.686	25 612.159	12.319	.911
$H_2 : \beta_{supervision} = \beta_{same\ team} = \beta_{same\ affiliation} > 0$	0.017	1.000	429.117	0.206	.015
$H_3 : \beta_{supervision} = \beta_{same\ team} = \beta_{same\ affiliation} = 0$	0.000	0.002	1.000	0.000	.000
$H_4 : complement$	0.081	4.845	2079.047	1.000	.074

(both having large posterior standard deviations), which conflicts with the constraint $\beta_{supervision} > \beta_{same\ team}$ under H_1 . The clear support in favor of H_1 , however, can be explained by the fact that a very slight violation of one order constraint (with a lot of uncertainty) can be compensated by clear support for the other order constraints. We can also see this when looking at the posterior and prior probability that the constraints of H_1 hold under the full unconstrained ERGM, i.e., the numerator and denominator of the second term on the right hand side of Eq. (10), which are equal to 0.460 and .064, respectively.⁷ Thus, the relative fit is much larger than the relative complexity of H_1 resulting in clear support for the order-constrained hypothesis (see also Hoijtink et al., 2019).

6. Discussion

A Bayes factor was proposed for testing a broad class of statistical hypotheses with competing equality and/or order constraints on the coefficients under ERGMs. The outcome of the Bayes factor quantifies the relative evidence in the data in favor of a hypothesis with a set of constraints on the ERGM coefficients relative to a hypothesis with alternative constraints on the coefficients. When prior probabilities for the

hypotheses are formulated, the Bayes factors can be translated to posterior probabilities of the hypotheses, which quantify the plausibility of the hypothesis after observing the data. Several attractive properties of the methodology were illustrated, such as its consistent behavior, its ability to test multiple hypotheses simultaneously, its ability to quantify the evidence in the data in favor of the null, and its ability to balance between fit and complexity as an Occam’s razor, also for hypotheses with order constraints. These properties are not shared by classical p -values (which are currently dominant for hypothesis testing in applied social network research using ERGMs). Consequently, when observing a nonsignificant testing outcome, this does not tell us whether this implies there is evidence of absence of an effect (i.e., support for the simpler model) or absence of evidence (i.e., inconclusive evidence). Bayes factors, on the other hand, are able to distinguish between these two scenarios.

The proposed Bayes factors were based on a g -prior that contains the information of one dyad (Zellner, 1986; Liang et al., 2008). A flat prior was specified for the edges parameter (which serves as an intercept) together with a multivariate normal prior for the ERGM coefficients having zero means and a prior covariance structure that was based on the design matrix of the change statistics of the full ERGM. Zero prior means imply that the prior is neutral regarding the direction of the effect. Thereby the prior differs from the unit-information on which the BIC is based which is centered at the MLE (Schwarz, 1978; Raftery, 1995). When testing simple null hypotheses of whether an effect equals zero, the proposed Bayes factor was less conservative towards the null

⁷ These values are printed in the Specification table when calling summary() of the BF object (see also Mulder et al., 2021b).

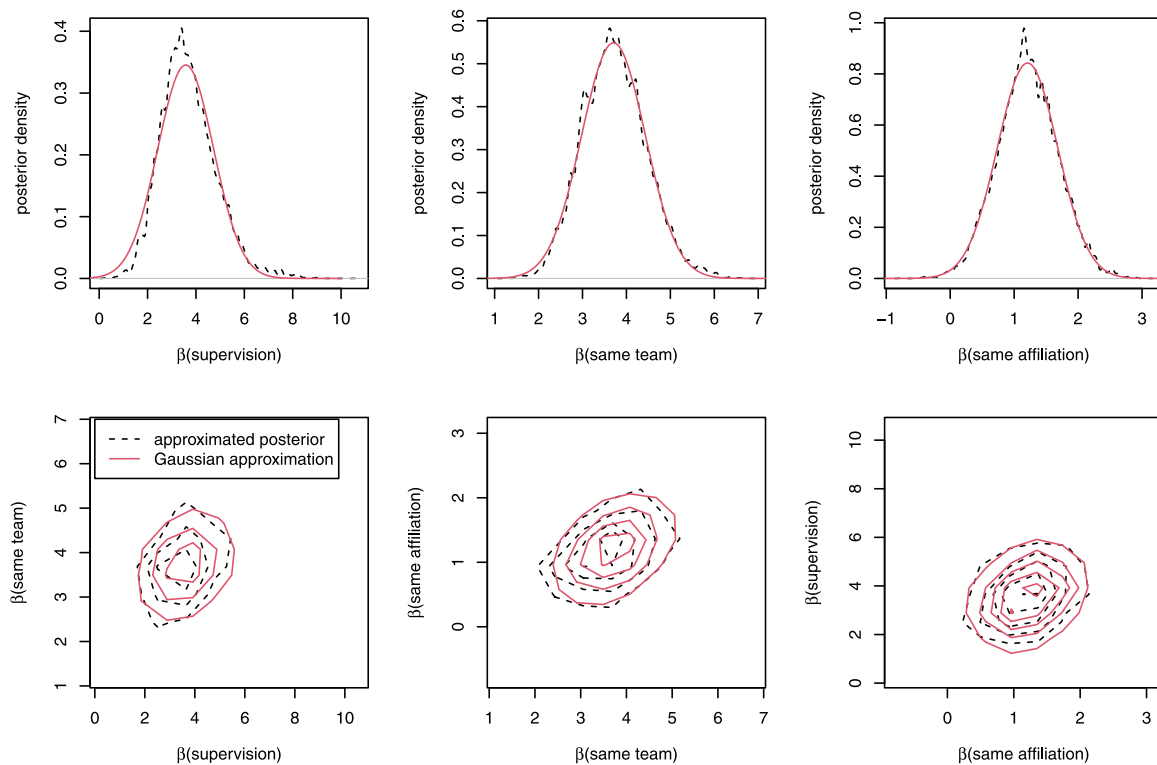


Fig. 7. Posterior density estimate based on 1,000,000 posterior draws (dashed black line) and the Gaussian approximation of the posterior (solid red line) for the co-authorship network.

Table 4

Classical results (MLE, standard error, and p -value) and posterior probabilities of a null, a negative, and a positive effect (assuming equal prior probabilities) for all ERGM coefficients of the co-authorship network (Application B).

	Classical results			Bayesian estimates		Posterior probabilities		
	MLE	s.e.	p	post. mean	post. sd	$P(\beta = 0 \mathbf{Y})$	$P(\beta < 0 \mathbf{Y})$	$P(\beta > 0 \mathbf{Y})$
edges	-10.099	0.625	0.000	-9.010	0.842	-	-	-
supervision	2.864	0.367	0.000	3.692	1.103	0.035	0.001	0.965
same team	2.594	0.298	0.000	3.769	0.681	0.000	0.000	1.000
same affiliation	1.052	0.229	0.000	1.210	0.459	0.109	0.004	0.887
GWESP	1.507	0.164	0.000	1.552	0.240	0.000	0.000	1.000
GWDegree	0.983	0.286	0.001	0.406	0.385	0.569	0.061	0.370
seniority	0.125	0.120	0.297	0.288	0.271	0.630	0.054	0.316
gender	0.354	0.184	0.053	-0.031	0.297	0.727	0.151	0.122
gender homophily	0.042	0.234	0.858	0.192	0.393	0.710	0.092	0.197
geographic distance	-0.239	0.150	0.111	-0.502	0.288	0.350	0.622	0.028
topic similarity	15.476	2.120	0.000	16.782	4.590	0.005	0.000	0.994
share Eng. articles	0.407	0.184	0.027	0.263	0.419	0.708	0.077	0.215
number of publications	0.022	0.005	0.000	0.019	0.011	0.428	0.024	0.547
Eng. art. similarity	2.038	0.495	0.000	1.432	0.857	0.350	0.032	0.618

in comparison to the BIC (which is sometimes a critique of the BIC, e.g., see Weakliem, 1999). The AIC, on the other hand, was more liberal than the proposed Bayes factor, which can be explained by the fact that the AIC is not consistent as a measure of evidence if a null model is true (O’Hagan, 1995).

To test equality/order-constrained hypotheses, truncated versions of the full unconstrained prior were constructed, so that the Bayes factors between any pair of hypotheses can be computed extremely fast using extended Savage–Dickey ratios and Gaussian approximations of the (kernel of the) posterior. This prior approach has also been considered by Klugkist and Hoijtink (2007) for testing constrained hypotheses, for instance. Furthermore, these truncated priors naturally appear in partial Bayes factors and fractional Bayes factors (Berger and Pericchi, 1996; O’Hagan, 1995; Mulder et al., 2010; Mulder, 2014; Mulder and Gu, 2022). For interesting discussions of this approach and possible alternatives (in the realm of linear models), we refer the interested reader to Consonni and Veronese (2008), Wetzels et al. (2010), Marin

and Robert (2010), Heck (2019), Mulder et al. (2021a) and Mulder et al. (2022), and the references therein.

Furthermore, the two empirical applications showed some differences between the classical estimates and the Bayesian estimates based on the proposed unit-information prior. Even though one might expect that the Bayesian estimates are closer to zero (due to prior shrinkage towards the prior mean of zero), some Bayesian estimates and posterior standard deviations were larger than their classical counterparts. We also observed this when fitting a Bayesian ERGM using flat priors. Therefore the observed differences between the Bayesian and classical estimates are not necessarily attributed to the proposed prior. In summary, prior specification under ERGMs remains an important topic for Bayesian hypothesis testing and Bayesian estimation problems, which deserves further attention in the literature.

Finally, the implementation of the new methodology in the R package BFpack (Mulder et al., 2021b) allows researchers to apply the

methodology in a straightforward manner. Thereby, the paper contributes to the increasing availability of Bayes factors for social network analysis, such as network autocorrelation models (Dittrich et al., 2019, 2020), relational event models (Mulder and Leenders, 2019), as well as other statistical frameworks such as structural equation models (Gu et al., 2019; Van Lissa et al., 2021), or (multivariate) regression models (Rouder and Morey, 2012; Braeken et al., 2015; Mulder and Gu, 2022). For a recent overview of applications of Bayes factor testing in the social and behavioral sciences, we refer interested readers to Heck et al. (2022). Furthermore, tutorials on the use of Bayes factors are available by Masson (2011) and Wagenmakers et al. (2018), or Hoijtink et al. (2019).

Acknowledgments

The authors wish to thank two anonymous reviewers for their constructive feedback which helped to improve the manuscript. The first author was supported by an ERC Starting Grant (758791).

Appendix. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.socnet.2023.11.004>.

References

- Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.-J., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., et al., 2018. Redefine statistical significance. *Nat. Hum. Behav.* 2 (1), 6–10.
- Berger, J.O., Pericchi, L.R., 1996. The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* 91, 109–122.
- Bouranis, L., Friel, N., Maire, F., 2018. Bayesian model selection for exponential random graph models via adjusted pseudolikelihoods. *J. Comput. Graph. Statist.* 27 (3), 516–528.
- Braeken, J., Mulder, J., Wood, S., 2015. Relative effects at work: Bayes factors for order hypotheses. *J. Manage.* 41 (2), 544–573.
- Breiger, R., Pattison, P., 1986. Cumulated social roles: The duality of persons and their algebras. *Social Networks* 8 (2), 215–256.
- Caimo, A., Friel, N., 2013. Bayesian model selection for exponential random graph models. *Social Networks* 35 (1), 11–24.
- Caimo, A., Friel, N., 2014. Bergm: Bayesian exponential random graphs in R. *J. Stat. Softw.* 61 (2), 1–25.
- Consonni, G., Veronese, P., 2008. Compatibility of prior specifications across linear models. *Statist. Sci.* (3), 332–353.
- Dickey, J., 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Statist.* 42, 204–223.
- Dienes, Z., 2014. Using Bayes to get the most out of non-significant results. *Front. Psychol.* 5, 781.
- Dittrich, D., Leenders, R.T.A., Mulder, J., 2019. Network autocorrelation modeling: A Bayes factor approach for testing (multiple) precise and interval hypotheses. *Sociol. Methods Res.* 48 (3), 642–676.
- Dittrich, D., Leenders, R.T.A., Mulder, J., 2020. Network autocorrelation modeling: Bayesian techniques for estimating and testing multiple network autocorrelations. *Sociol. Methodol.* 50 (1), 168–214.
- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis.
- Goodreau, S.M., Kitts, J.A., Morris, M., 2009. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* 46 (1), 103–125. <http://dx.doi.org/10.1353/dem.0.0045>.
- Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* 31 (4), 337–350.
- Gu, X., Hoijtink, H., Mulder, J., Rosseel, Y., 2019. Bain: A program for the evaluation of inequality constrained hypotheses using Bayes factors in structural equation models. *J. Stat. Comput. Simul.*
- Hanson, T.E., Branscum, A.J., Johnson, W.O., 2014. Informative g-priors for logistic regression. *Bayesian Anal.* 9 (3), 597–612.
- Heck, D.W., 2019. A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *Br. J. Math. Stat. Psychol.* 72 (2), 316–333.
- Heck, D.W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H.A., et al., 2022. A review of applications of the Bayes factor in psychological research. *Psychol. Methods* <http://dx.doi.org/10.1037/met0000454>.
- Henry, A.D., Lubell, M., McCoy, M., 2011. Belief systems and social capital as drivers of policy network structure: The case of California regional planning. *J. Public Adm. Res. Theory* 21 (3), 419–444. <http://dx.doi.org/10.1093/jopart/muq042>.
- Hoijtink, H., 2011. Informative Hypotheses: Theory and Practice for Behavioral and Social Scientists. Chapman & Hall/CRC, New York.
- Hoijtink, H., Mulder, J., van Lissa, C., Gu, X., 2019. A tutorial on testing hypotheses using the Bayes factor. *Psychol. Methods* 24 (5), 539.
- Hunter, D.R., Handcock, M.S., Butts, C.T., Goodreau, S.M., Morris, M., 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* 24 (3), 1–29.
- Jeffreys, H., 1961. *Theory of Probability*-3rd Ed. Oxford University Press, New York.
- Kass, R.E., Wasserman, L., 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Statist. Assoc.* 90 (431), 928–934.
- Kim, J.Y., Howard, M., Cox Pahnke, E., Boeker, W., 2016. Understanding network formation in strategy research: Exponential random graph models. *Strateg. Manag. J.* 37 (1), 22–44. <http://dx.doi.org/10.1002/smj.2454>.
- Klugkist, I., Hoijtink, H., 2007. The Bayes factor for inequality and about equality constrained models. *Comput. Statist. Data Anal.* 51, 6367–6379.
- Krivitsky, P.N., Hunter, D.R., Morris, M., Klumb, C., 2023. Ergm 4: New features for analyzing exponential-family random graph models. *J. Stat. Softw.* 105, 1–44.
- Leifeld, P., 2018. Polarization in the social sciences: Assortative mixing in social science collaboration networks is resilient to interventions. *Physica A* 507, 510–523. <http://dx.doi.org/10.1016/j.physa.2018.05.109>.
- Leifeld, P., Schneider, V., 2012. Information exchange in policy networks. *Am. J. Political Sci.* 56 (3), 731–744. <http://dx.doi.org/10.1111/j.1540-5907.2011.00580.x>.
- Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2008. Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* 103 (481), 410–423.
- Lindley, D.V., 1957. A statistical paradox. *Biometrika* 44, 187–192.
- Marin, J.-M., Robert, C.P., 2010. On resolving the Savage–Dickey paradox. *Electron. J. Stat.* 4, 643–654.
- Masson, M., 2011. A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behav. Res. Methods* 43, 679–690.
- Mulder, J., 2014. Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Comput. Statist. Data Anal.* 71, 448–463. <http://dx.doi.org/10.1016/j.csa.2013.07.017>.
- Mulder, J., Berger, J.O., Peña, V., Bayarri, M., 2021a. On the prevalence of information inconsistency in normal linear models. *Test* 30, 103–132. <http://dx.doi.org/10.1007/s11749-020-00704-4>.
- Mulder, J., Gelissen, J.P., 2021. Bayes factor testing of equality and order constraints on measures of association in social research. *J. Appl. Stat.* 1–37. <http://dx.doi.org/10.1080/02664763.2021.1992360>.
- Mulder, J., Gu, X., 2022. Bayesian testing of scientific expectations under multivariate normal linear models. *Multivar. Behav. Res.* 57 (5), 767–783. <http://dx.doi.org/10.1080/00273171.2021.1904809>.
- Mulder, J., Hoijtink, H., Klugkist, I., 2010. Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *J. Statist. Plann. Inference* 140, 887–906. <http://dx.doi.org/10.1016/j.jspi.2009.09.022>.
- Mulder, J., Leenders, R.T.A., 2019. Modeling the evolution of interaction behavior in social networks: A dynamic relational event approach for real-time analysis. *Chaos Solitons Fractals* 119, 73–85. <http://dx.doi.org/10.1016/j.chaos.2018.11.027>.
- Mulder, J., Wagenmakers, E.-J., 2016. Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *J. Math. Psych.* 72, 1–5.
- Mulder, J., Wagenmakers, E.-J., Marsman, M., 2022. A generalization of the Savage–Dickey density ratio for testing equality and order constrained hypotheses. *Amer. Statist.* 76 (2), 102–109. <http://dx.doi.org/10.1080/00031305.2020.1799861>.
- Mulder, J., Williams, D., Gu, X., Tomarken, A., Böing-Messing, F., Olsson-Collentine, A., Meijerink, M., Menke, J., van Aert, R., Fox, J.-P., et al., 2021b. BFpack: Flexible Bayes factor testing of scientific theories in R. *J. Stat. Softw.* 100, <http://dx.doi.org/10.18637/jss.v100.i18>.
- O'Hagan, A., 1995. Fractional Bayes factors for model comparison (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57, 99–138. <http://dx.doi.org/10.1111/j.2517-6161.1995.tb02017.x>.
- Raftery, A.E., 1995. Bayesian model selection in social research. *Sociol. Methodol.* 25, 111–163. <http://dx.doi.org/10.2307/271063>.
- Raftery, A.E., 1999. Bayes factors and BIC: Comment on "a critique of the Bayesian information criterion for model selection". *Sociol. Methods Res.* 27 (3), 411–427.
- Rouder, J.N., Morey, R.D., 2012. Default Bayes factors for model selection in regression. *Multivar. Behav. Res.* 47 (6), 877–903.
- Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G., 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* 16, 225–237.
- Saul, Z.M., Filkov, V., 2007. Exploring biological network structure using exponential random graph models. *Bioinformatics* 23 (19), 2604–2611. <http://dx.doi.org/10.1093/bioinformatics/btm370>.
- Schwarz, G.E., 1978. Estimating the dimension of a model. *Ann. Statist.* 6, 461–464.
- Silvapulle, M.J., Sen, P.K., 2004. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions*, second ed. John Wiley, Hoboken, NJ.

- Simpson, S.L., Hayasaka, S., Laurienti, P.J., 2011. Exponential random graph modeling for complex brain networks. *PloS ONE* 6 (5), e20039. <http://dx.doi.org/10.1371/journal.pone.0020039>.
- Van Lissa, C.J., Gu, X., Mulder, J., Rosseel, Y., Van Zundert, C., Hoijtink, H., 2021. Teacher's corner: Evaluating informative hypotheses using the Bayes factor in structural equation models. *Struct. Equ. Model.: Multidiscip. J.* 28 (2), 292–301.
- Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., Grasman, R., 2010. Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cogn. Psychol.* 60, 158–189.
- Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q.F., Šmíra, M., Epskamp, S., et al., 2018. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon. Bull. Rev.* 25, 35–57. <http://dx.doi.org/10.3758/s13423-017-1343-3>.
- Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p. *Psychometrika* 61 (3), 401–425.
- Weakliem, D.L., 1999. A critique of the Bayesian information criterion for model selection. *Sociol. Methods Res.* 27 (3), 359–397.
- Wetzels, R., Grasman, R.P.P.P., Wagenmakers, E.J., 2010. An encompassing prior generalization of the Savage–Dickey density ratio test. *Comput. Statist. Data Anal.* 38, 666–690.
- Yin, F., Butts, C.T., 2022. Highly scalable maximum likelihood and conjugate Bayesian inference for ERGMs on graph sets with equivalent vertices. *Plos one* 17 (8), e0273039.
- Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In: Goel, P.K., Zellner, A. (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. Amsterdam: North-Holland/Elsevier, pp. 233–243.