A Hypotheses-driven Bayesian Approach for Understanding Edge Formation in Attributed Multigraphs

Lisette Espín-Noboa,Florian Lemmerich, Markus Strohmaier and Philipp Singer

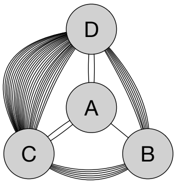
Abstract Understanding edge formation represents a key question in network analysis. Various approaches have been postulated across disciplines ranging from network growth models to statistical (regression) methods. In this work, we extend this existing arsenal of methods with a hypotheses-driven Bayesian approach that allows to intuitively compare hypotheses about edge formation on attributed multigraphs. We model the multiplicity of edges using a simple categorical model and propose to express hypotheses as priors encoding our belief about parameters. Using Bayesian model comparison techniques, we compare the relative plausibility of hypotheses which might be motivated by previous theories about edge formation based on popularity or similarity. We demonstrate the utility of our approach on synthetic and empirical data. This work is relevant for researchers interested in studying mechanisms explaining edge formation in networks.

1 Introduction

Understanding edge formation in networks is a key interest of our research community. For example, social scientists are frequently interested in studying relations between entities within social networks, e.g., how social friendship ties form between actors and explain them based on attributes such as a person's gender, race, political affiliation or age in the network [18]. Similarly, the complex networks community suggests a set of generative network models aiming at explaining the formation of edges focusing on the two core principles of *popularity* and *similarity* [15]. Thus, a series of approaches to study edge formation have emerged including statistical (regression) tools [10, 23] and model-based approaches [6, 15, 24] specifically established in the physics and complex networks communities. Other disciplines such as the computer sciences, biomedical sciences or political sciences use these tools to

Lisette Espín-Noboa✉ · Florian Lemmerich✉ · Markus Strohmaier✉ · Philipp Singer✉
GESIS, University of Koblenz-Landau, e-mail: `firstname.(first)lastname@gesis.org`

H. Cherifi et al. (eds.), *Complex Networks & Their Applications V*,
Studies in Computational Intelligence 693,
DOI 10.1007/978-3-319-50901-3_2



(a) Multigraph

	A	B	C	D
A	0	1	2	2
B	1	0	6	7
C	2	6	0	26
D	2	7	26	0

(b) Adjacency Matrix

attribute/node	A	B	C	D
country	Ecuador	Germany	Austria	Austria
gender	F	M	M	M
position	1	2	2	3
academic	2015	2008	2011	2001
articles	3	37	30	115
citations	17	280	203	1918

(c) Node Attributes

Fig. 1: **Example**: This example illustrates an unweighted attributed multigraph. (a) Shows a multigraph where nodes represent academic researchers, and edges scientific articles in which they have collaborated together. (b) Shows the adjacency matrix of the graph, where every cell represents the total number of edges between two nodes. (c) Decodes some attribute values per node. For instance, node D shows information about an *Austrian* researcher who started *his* academic career in *2001*.

answer empirical questions; e.g., co-authorship networks[12], wireless networks of biomedical sensors [19], or community structures of political blogs [1].

Problem Illustration. For illustration, consider Fig. 1; nodes represent authors, and (multiple) edges between them refer to co-authored scientific articles. Node attributes provide additional information about the authors, e.g., their home country and gender. An exemplary research question could be: “Can co-authorship be better explained by a mechanism that assumes more collaborations between authors from the *same country* or by a mechanism that assumes more collaborations between authors with the *same gender*?”. These and similar questions motivate the main objective of this work, which is to provide a Bayesian approach for understanding how edges emerge in networks based on some characteristics of the nodes.

While several methods for tackling such questions have been proposed, they come with certain limitations. For example, statistical regression methods based on QAP [5] or mixed-effects models [20] do not scale to large-scale data and results are difficult to interpret. For network growth models [15], it is necessary to find the appropriate model for a given hypothesis about edge formation and thus, it is often not trivial to intuitively compare competing hypotheses that sometimes might even go beyond simple popularity and similarity mechanisms. Consequently, we want to extend the methodological toolbox for studying edge formation in networks by proposing a first step towards a hypotheses-driven generative Bayesian framework.

Approach and methods. We focus on understanding edge formation in node-attributed multigraphs. We are interested in modeling and understanding the multiplicity of edges based on node attributes. Our approach follows a generative storyline. First, we define the model that can characterize the edge formation at interest. We focus on the simple categorical model, from which edges are independently drawn from. Motivated by previous work on sequential data [21], the core idea of our approach is to specify generative hypotheses about how edges emerge in a network. These hypotheses might be motivated by previous theories such as popularity or similarity [15]—e.g., for Fig. 1 we could hypothesize that authors are more likely to collaborate with each other if they are from the same country. Technically, we elicit these types of hypotheses as beliefs in parameters of the underlying categorical model and encode and integrate them as priors into the Bayesian framework. Using Bayes factors with marginal likelihood estimations allows us to compare the relative plausibility of expressed hypotheses as they are specifically sensitive to the priors. The final output is a ranking of hypotheses based on their plausibility given the data.

Contributions. Our main contributions are: (i) We present a first step towards a Bayesian approach for comparing generative hypotheses about edge formation in networks. (ii) We provide simple categorical models based on local and global scenarios allowing the comparison of hypotheses for multigraphs. (iii) We provide guidelines for building hypotheses based on node attributes. (iv) We demonstrate the applicability of our approach on synthetic and empirical data. (v) We make an implementation of this approach openly available[1] on the Web.

2 Background

We start by introducing the underlying concepts of our approach.

Attributed Multigraphs. In this paper, we focus on *multigraphs* with *attributed nodes* and *unweighted edges without own identity*. That means, each pair of nodes can be connected by multiple indistinguishable edges, and there are features for the individual nodes available.

We formally define this as: Let $G = (V, E, F)$ be an unweighted attributed multigraph with $V = (v_1, \dots, v_n)$ being a list of nodes, $E = \{(v_i, v_j)\} \in V \times V$ a multiset of either directed or undirected edges, and a set of feature vectors $F = (f_1, \dots, f_n)$. Each feature vector $f_i = (f_i[1], ..., f_i[c])^T$ maps a node v_i to c (numeric or categorical) attribute values. The graph structure is captured by an adjacency matrix $M_{n\times n} = (m_{ij})$, where m_{ij} is the multiplicity of edge (v_i, v_j) in E (i.e., number of edges between nodes v_i and v_j). By definition, the total number of multiedges is $l = |E| = \sum_{ij} m_{ij}$.

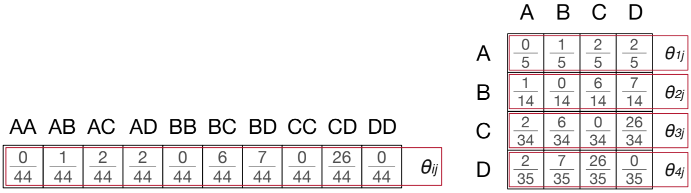
Fig. 1a shows an example unweighted attributed multigraph: nodes represent authors, and undirected edges represent co-authorship in scientific articles. The adjacency matrix of this graph—counting for multiplicity of edges—is shown in Fig. 1b. Feature vectors (node attributes) are described in Fig. 1c. Thus, for this particular case, we account for $n = 4$ nodes, $l = 44$ multiedges, and $c = 6$ attributes.

Bayesian Hypothesis Testing. Our approach compares hypotheses on edge formation based on techniques from Bayesian hypothesis testing [11, 21]. The elementary Bayes' theorem states for parameters θ, given data D and a hypothesis H that:

$$\overbrace{P(\theta|D,H)}^{\text{posterior}} = \frac{\overbrace{P(D|\theta,H)}^{\text{likelihood}}\overbrace{P(\theta|H)}^{\text{prior}}}{\underbrace{P(D|H)}_{\text{marginal likelihood}}} \tag{1}$$

As observed data D, we use the adjacency matrix M, which encodes edges counts. θ refers to the model parameters, which in our scenario correspond to the probabilities of individual edges. H denotes a hypothesis under investigation. The *likelihood*

[1] https://github.com/lisette-espin/JANUS



AA	AB	AC	AD	BB	BC	BD	CC	CD	DD	
$\frac{0}{44}$	$\frac{1}{44}$	$\frac{2}{44}$	$\frac{2}{44}$	$\frac{0}{44}$	$\frac{6}{44}$	$\frac{7}{44}$	$\frac{0}{44}$	$\frac{26}{44}$	$\frac{0}{44}$	θ_{ij}

(a) Global

	A	B	C	D	
A	$\frac{0}{5}$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{2}{5}$	θ_{1j}
B	$\frac{1}{14}$	$\frac{0}{14}$	$\frac{6}{14}$	$\frac{7}{14}$	θ_{2j}
C	$\frac{2}{34}$	$\frac{6}{34}$	$\frac{0}{34}$	$\frac{26}{34}$	θ_{3j}
D	$\frac{2}{35}$	$\frac{7}{35}$	$\frac{26}{35}$	$\frac{0}{35}$	θ_{4j}

(b) Local

Fig. 2: **Multigraph models**: This figure shows two ways of modeling the undirected multigraph shown in Fig. 1. That is, (a) global or graph-based model models the whole graph as a single distribution. (b) Local or neighbour-based model models each node as a separate distribution.

describes, how likely we observe data D given parameters θ and a hypothesis H. The *prior* is the distribution of parameters we believe in before seeing the data; in other words, the prior encodes our hypothesis H. The *posterior* represents an adjusted distribution of parameters after we observe D. Finally, the *marginal likelihood* (also called *evidence*) represents the probability of the data D given a hypothesis H.

In our approach, we exploit the sensitivity of the marginal likelihood on the prior to compare and rank different hypotheses: more plausible hypotheses imply higher evidence for data D. Formally, *Bayes Factors* can be employed for comparing two hypotheses. These are computed as the ratio between the respective marginal likelihood score. The strength of a Bayes factor can be judged using available interpretation tables [7]. While in many cases determining the marginal likelihood is computationally challenging and requires approximate solutions, we can rely on exact and fast-to-compute solutions in the models employed in this paper.

3 Approach

In this section, we describe the main steps towards a hypotheses-driven Bayesian approach for understanding edge formation in unweighted attributed multigraphs. To that end, we propose intuitive models for edge formation (Section 3.1), a flexible toolbox to formally specify belief in the model parameters (Section 3.2), a way of computing proper (Dirichlet) priors from these beliefs (Section 3.2), computation of the marginal likelihood in this scenario (Section 3.3), and guidelines on how to interpret the results (Section 3.4). We subsequently discuss these issues one-by-one.

3.1 Generative Edge Formation Models

We propose two variations of our approach, which employ two different types of generative edge formation models in multigraphs.

Global model. First, we utilize a simple *global model*, in which a fixed number of graph edges are randomly and independently drawn from the set of all potential edges in the graph G by sampling with replacement. Each edge (v_i, v_j) is sampled from

	A	B	C	D
A	0.0	0.1	0.1	0.1
B	0.1	0.0	0.1	0.1
C	0.1	0.1	0.0	0.9
D	0.1	0.1	0.9	0.0

(a) Belief matrix B_1

	A	B	C	D
A	0.00	0.33	0.33	0.33
B	0.33	0.00	0.33	0.33
C	0.09	0.09	0.00	0.82
D	0.09	0.09	0.82	0.00

(b) Normalized B_1

	A	B	C	D
A	1.00	2.33	2.33	2.33
B	2.33	1.00	2.33	2.33
C	1.36	1.36	1.00	4.27
D	1.36	1.36	4.27	1.00

(c) Prior $\kappa = 4$

Fig. 3: **Prior belief**: This figure illustrates the three main phases of prior elicitation. That is, (a) a matrix representation of belief B_1, where authors are more likely to collaborate with each other if they are from the same country. (b) B_1 normalized row-wise using the local model interpretation. (c) Prior elicitation for $\kappa = 4$; i.e., $\alpha_{ij} = \frac{b_{ij}}{Z} \times \kappa + 1$.

a *categorical distribution* with parameters $\theta_{ij}, 1 \leq i \leq n, 1 \leq j \leq n, \forall ij : \sum_{ij} \theta_{ij} = 1$: $(v_i, v_j) \sim Categorical(\theta_{ij})$. This means that each edge is associated with one probability θ_{ij} of being drawn next. Fig. 2a shows the maximum likelihood global model for the network shown in Fig. 1. Since this is an undirected graph, inverse edges can be ignored resulting in $n(n+1)/2$ potential edges/parameters.

Local models. As an alternative, we can also focus on a *local level*. Here, we model to which other node a specific node v will connect *given that any new edge starting from* v is formed. We implement this by using a set of n separate models for the outgoing edges of the ego-networks (i.e., the 1-hop neighborhood) of each of the n nodes. The ego-network model for node v_i is built by drawing randomly and independently a number of nodes v_j by sampling with replacement and adding an edge from v_i to this node. Each node v_j is sampled from a *categorical distribution* with parameters $\theta_{ij}, 1 \leq i \leq n, 1 \leq j \leq n, \forall i : \sum_j \theta_{ij} = 1$: $v_j \sim Categorical(\theta_{ij})$. The parameters θ_{ij} can be written as a matrix; the value in cell (i, j) specifies the probability that a new formed edge with source node v_i will have the destination node v_j. Thus, all values within one row always sum up to one. Local models can be applied for undirected and directed graphs (cf. also discussion in Section 6). In the directed case, we model only the outgoing edges of the ego-network. Fig. 2b depicts the maximum likelihood local models for our introductory example .

3.2 Hypothesis Elicitation

The main idea of our approach is to encode our beliefs in edge formation as Bayesian priors over the model parameters. As a common choice, we employ Dirichlet distributions as the *conjugate priors* of the categorical distribution. Thus, we assume that the model parameters θ are drawn from a Dirichlet distribution with hyperparameters α: $\theta \sim Dir(\alpha)$. Similar to the model parameters themselves, the Dirichlet prior (or multiple priors for the local models) can be specified in a matrix. We will choose the parameters α in such a way that they reflect a specific belief about edge formation. For that purpose, we first specify matrices that formalize these beliefs, then we compute the Dirichlet parameters α from these beliefs.

Constructing Belief Matrices. We specify hypotheses about edge formation as *belief matrices* $B = b_{ij}$. These are $n \times n$ matrices, in which each cell $b_{ij} \in \mathbb{R}$ represents a belief of having an edge from node v_i to node v_j. To express a belief that an edge occurs more often (compared to other edges) we set b_{ij} to a higher value. In general, users have a large freedom to generate belief matrices. However, typical construction principles are to assume that nodes with specific attributes are more *popular* and thus edges connecting these attributes receive higher multiplicity, or to assume that nodes that are *similar* with respect to one or more attributes are more likely to form an edge, cf. [15]. Ideally, the elicitation of belief matrices is based on existing theories.

For example, based on the information shown in Fig. 1, one could "believe" that two authors collaborate *more frequently* together if: (1) they both are from the same country, (2) they share the same gender, (3) they have high positions, or (4) they are popular in terms of number of articles and citations. We capture each of these beliefs in one matrix. One implementation of the matrices for our example beliefs could be:

- B_1 (same country): $b_{ij} := 0.9$ if $f_i[country] = f_j[country]$ and 0.1 otherwise
- B_2 (same gender): $b_{ij} := 0.9$ if $f_i[gender] = f_j[gender]$ and 0.1 otherwise
- B_3 (hierarchy): $b_{ij} := f_i[position] \cdot f_j[position]$
- B_4 (popularity): $b_{ij} := f_i[articles] + f_j[articles] + f_i[citations] + f_j[citations]$

Fig. 3a shows the matrix representation of belief B_1, and Fig. 3b its respective row-wise normalization for the local model case. While belief matrices are identically structured for local and global models, the ratio between parameters in different rows is crucial for the global model, but irrelevant for local ones.

Eliciting a Dirichlet prior. In order to obtain the hyperparameters α of a prior Dirichlet distribution we utilize the pseudo-count interpretation of the parameters α_{ij} of the Dirichlet distribution, i.e., a value of α_{ij} can be interpreted as $\alpha_{ij} - 1$ previous observations of the respective event for $\alpha_{ij} \geq 1$. We distribute pseudo-counts proportionally to a belief matrix. Consequently, the hyperparameters can be expressed as: $\alpha_{ij} = \frac{b_{ij}}{Z} \times \kappa + 1$, where κ is the concentration parameter of the prior. The normalization constant Z is computed as the sum of all entries of the belief matrix in the global model, and as the respective row sum in the local case. We suggest to set $\kappa = n \times k$, $k = \{0, 1, ..., 10\}$. A high value of κ expresses a strong belief in the prior parameters. A similar alternative method to obtain Dirichlet priors is the *trial roulette method* [21]. For the global model variation, all α values are parameters for the same Dirichlet distribution, whereas in the local model variation, each row parametrizes a separate Dirichlet distribution.

3.3 Computation of the Marginal Likelihood

For comparing the relative plausibility of hypotheses we use the marginal likelihood. This is the aggregated likelihood over all possible values of the parameters θ weighted by the Dirichlet prior. For our set of local models we can calculate them as:

$$P(D|H) = \prod_{i=1}^{n} \frac{\Gamma(\sum_{j=1}^{n} \alpha_{ij})}{\Gamma(\sum_{j=1}^{n} \alpha_{ij} + m_{ij})} \prod_{j=1}^{n} \frac{\Gamma(\alpha_{ij} + m_{ij})}{\Gamma(\alpha_{ij})} \tag{2}$$

Recall, α_{ij} encodes our prior belief connecting nodes v_i and v_j in G, and m_{ij} are the actual edge counts. Since we evaluate only a single model in the global case, the product over rows i of the adjacency matrix can be removed, and we obtain:

$$P(D|H) = \frac{\Gamma(\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij})}{\Gamma(\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_{ij} + m_{ij})} \prod_{i=1}^{n} \prod_{j=1}^{n} \frac{\Gamma(\alpha_{ij} + m_{ij})}{\Gamma(\alpha_{ij})} \tag{3}$$

Equation (3) holds for directed networks. In the undirected case, indices j go from i to n accounting for only half of the matrix including the diagonal. For a detailed derivation of the marginal likelihood given a Dirichlet-Categorical model see [22, 25]. For both models we focus on the log-marginal likelihoods in practice to avoid underflows.

Bayes Factor. Formally, we compare the relative plausibility of hypotheses by using so-called *Bayes factors* [7], which simply are the ratios of the marginal likelihoods for two hypotheses H_1 and H_2. If it is positive, the first hypothesis is judged as more plausible. The strength of the Bayes factor can be checked in an interpretation table provided by Kass and Raftery [7].

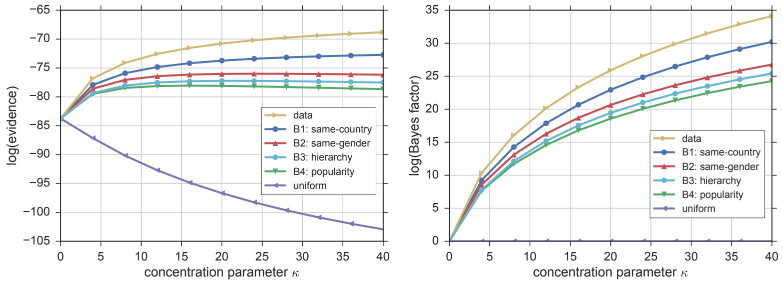
3.4 Application of the Method and Interpretation of Results

We now showcase an example application of our approach featuring the network shown in Fig. 1, and demonstrate how results can be interpreted. For that purpose and due to space limitations, we focus on the local models variant.

Hypotheses. We compare four hypotheses (represented as belief matrices) B_1, B_2, B_3, and B_4 elaborated in Section 3.2. Additionally, we use the *uniform hypothesis* as a *baseline*. It assumes that all edges are equally likely, i.e., $b_{ij} = 1$ for all i, j. Hypotheses that are not more plausible than the uniform cannot be assumed to capture relevant underlying mechanisms of edge formation. We also use the *data hypothesis* as an upper bound for comparison, which employs the observed adjacency matrix as belief: $b_{ij} = m_{ij}$.

Calculation and visualization. For each hypothesis H and every κ, we can elicit the Dirichlet priors (cf. Section 3.2), determine the aggregated marginal likelihood (cf. Section 3.3), and compare the plausibility of hypotheses compared to the uniform hypothesis at the same κ by calculating the logarithm of the Bayes factor as $log(P(D|H)) - log(P(D|H_{uniform}))$. We suggest two ways of visualizing the results, i.e., ploting the marginal likelihood values (Fig. 4a) or showing the Bayes factors (Fig. 4b) on the y-axis. In both cases, the x-axis refers to the concentration parameter κ. While the visualization showing directly the marginal likelihoods carries more information, visualizing Bayes factors makes it easier to spot smaller differences between the hypotheses.

Interpretation. Every line in both figures represents a hypothesis. In Fig. 4a, higher evidence values mean higher plausibility. Similarly, in Fig. 4b positive Bayes factors



(a) Evidences (b) Bayes Factors

Fig. 4: **Ranking of hypotheses for the introductory example**. Rankings can be visualized using (a) the marginal likelihood or evidence (y-axis), or (b) Bayes factors (y-axis) by setting the uniform hypothesis as a baseline to compare with; higher values refer to higher plausibility. The x-axis depicts the concentration parameter κ. For this example, authors from the multigraph shown in Fig. 1 appear to prefer to collaborate more often with researchers of the same country rather than due to popularity (i.e., number of articles and citations). Note that all hypotheses outperform the uniform, meaning that they all represent reasonable explanations of edge formation for the given graph.

mean that for a given κ, the hypothesis is judged to be more plausible than the uniform baseline hypothesis; here, the relative Bayes factors also provide a ranking. If evidences or Bayes factors are increasing with κ, we can interpret this as further evidence for the plausibility of expressed hypothesis as this means that the more we believe in it, the higher the Bayesian approach judges its plausibility. As a result for our example, we see that the hypothesis believing that two authors are more likely to collaborate if they are from the same country is the most plausible one (after the data hypothesis). In this example, all hypotheses appear to be more plausible than the baseline, but this is not necessarily the case in all applications.

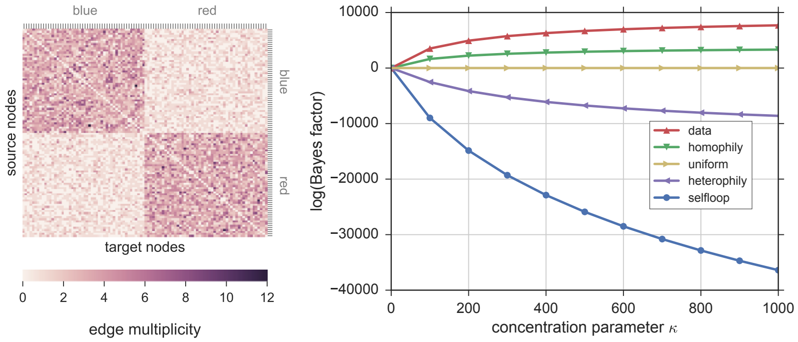
4 Experiments

We demonstrate the utility of our approach on both synthetic and empirical networks. Due to space limitations, we only showcase the local model results.

4.1 Synthetic Attributed Multigraph

We start with experiments on a synthetic attributed multigraph. Here, we control the underlying mechanisms of how edges in the network emerge and thus, expect these also to be good hypotheses for our approach.

Network. The network contains 100 nodes where each node is assigned one of two colors with uniform probability. For each node, we then randomly drew 200 undirected edges where each edge connects randomly with probability $p = 0.8$ to a



(a) Adjacency Matrix (b) Bayes Factors

Fig. 5: **Ranking of hypotheses for synthetic network.** In (a), we show the adjacency matrix of the 2-color random multigraph with a node correlation of 80% for nodes of the same color and 20% otherwise. One can see homophily based on more connections between nodes of the same color; the diagonal is zero as there are no self-connections. In (b), we show the ranking of hypotheses based on Bayes factors when compared to the uniform hypothesis. As expected, the homophily hypothesis explains the edge formation best (positive Bayes factor), and the heterophily and selfloop hypotheses show negative Bayes factors—i.e., they provide no good explanations for edge formation.

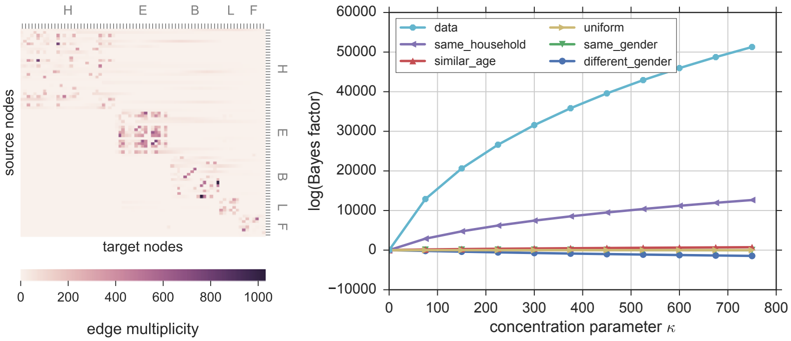
different node of the same color, and with $p = 0.2$ to a node of the opposite color. The adjacency matrix of this graph is visualized in Fig. 5a.

Hypotheses. In addition to the uniform baseline hypothesis, we construct two intuitive hypotheses based on the node color that express belief in possible edge formation mechanics. First, the *homophily hypothesis* assumes that nodes of the same color are more likely to have more edges between them. Therefore, we arbitrary set belief values b_{ij} to 80 when nodes v_i and v_j are of the same color, and 20 otherwise. Second, the *heterophily hypothesis* expresses the opposite behavior; i.e., $b_{ij} = 80$ if the color of nodes v_i and v_j are different, and 20 otherwise. An additional *selfloop hypothesis* only believes in self-connections (i.e., diagonal of adjacency matrix).

Results. Fig. 5b shows the ranking of hypotheses based on their Bayes factors compared to the uniform hypothesis. Clearly, the homophily hypothesis is judged as the most plausible. This is expected and corroborates the fact that network connections are biased towards nodes of the same color. The heterophily and selfloop hypotheses show negative Bayes factors; thus, they are not good hypotheses about edge formation in this network. Due to the fact that the multigraph lacks of selfloops, the selfloop hypothesis decreases very quickly with increasing strength of belief κ.

4.2 Empirical Attributed Multigraph

Here, we focus on a real-world contact network based on wearable sensors.



(a) Adjacency Matrix

(b) Bayes Factors

Fig. 6: **Ranking of hypotheses for Kenya contact network.** (a) Shows the adjacency matrix of the network with node ordering according to household membership. Darker cells indicate more contacts. (b) Displays the ranking of hypotheses based on Bayes factors, using the uniform hypothesis as baseline. The *same household hypothesis* (people are more likely to contact people from the same household) ranks highest. While the *similar age* hypothesis also provide positive Bayes Factors, the *same* and *different gender hypotheses* are less plausible than the baseline (uniform edge formation). Results are consistent for all κ.

Network. We study a network[2] capturing interactions of 5 households in rural Kenya between April 24 and May 12, 2012 [9]. The undirected unweighted multigraph contains 75 nodes (persons) and 32 643 multiedges (contacts) which we aim to explain. For each node, we know information such as gender and age (encoded into 5 age intervals). Interactions exist within and across households. Fig. 6a shows the adjacency matrix (i.e., number of contacts between two people) of the network. Household membership of nodes (rows/columns) is shown accordingly.

Hypotheses. We investigate edge formation by comparing—next to the uniform baseline hypothesis—four hypotheses based on node attributes as prior beliefs. (i) The *similar age hypothesis* expresses the belief that people of similar age are more likely to interact with each other. Entries b_{ij} of the belief matrix B are set to the inverse age distance between members: $\frac{1}{1+abs(f_i[age]-f_j[age])}$. (ii) The *same household hypothesis* believes that people are more likely to interact with people from the same household. We arbitrarily set b_{ij} to 80 if person v_i and person v_j belong to the same household, and 20 otherwise. (iii) With the *same gender hypothesis* we hypothesize that the number of same-gender interactions is higher than the different-gender interactions. Therefore, every entry b_{ij} of B is set to 80 if persons v_i and v_j are of the same gender, and 20 otherwise. Finally, (iv) the *different gender hypothesis* believes that it is more likely to find different-gender than same-gender interactions; b_{ij} is set to 80 if person v_i has the opposite gender of person v_j, and 20 otherwise.

Results. The results shown in Fig. 6b indicate that the *same household hypothesis* explains the data the best, since it has been ranked first and it is more plausible than the uniform. The *similar age* hypothesis also indicates plausibility due to positive

[2] http://www.sociopatterns.org/datasets/kenyan-households-contact-network/

Bayes factors. Both the *same* and *different gender hypotheses* show negative Bayes factors when compared to the uniform hypothesis suggesting that they are not good explanations of edge formation in this network. This gives us a better understanding of potential mechanisms producing underlying edges. People prefer to contact people from the same household and similar age, but not based on gender preferences. Additional experiments could further refine these hypotheses (e.g., combining them).

5 Related Work

We provide a broad overview of research on modeling and understanding edge formation in networks; i.e., *edge formation models* and *hypothesis testing on networks*.

Edge formation models. A variety of models explaining underlying mechanisms of *network formation* have been proposed. Here, we focus on models explaining linkage between dyads beyond structure by incorporating node attribute information. Prominently, the *stochastic blockmodel* [6] aims at producing and explaining communities by accounting for node correlation based on attributes. The *attributed graph* [16] models network structure and node attributes by learning the attribute correlations in the observed network. Furthermore, the *multiplicative attributed graph* [8] takes into account attribute information from nodes to model network structure. This model defines the probability of an edge as the product of individual attribute link formation affinities. *Exponential random graph models* [17] (also called the p^* class of models) represent graph distributions with an exponential linear model that uses feature-structure counts such as reciprocity, k-stars and k-paths. In this line of research, *p1 models* [4] consider expansiveness (sender) and popularity (receiver) as fixed effects associated with unique nodes in the network [3], in contrast to the *p2 models* [17] which account for random effects and assume dyadic independence conditionally to node-level attributes. While many of these works focus on binary relationships, [27] proposes an unsupervised model to estimate continuous-valued relationship strength for links from interaction activity and user similarity in social networks.

Hypothesis testing on networks. Previous works have implemented different techniques to test hypotheses about network structure. For instance, the work in [13] proposes an algorithm to determine whether two observed networks are significantly different. Another branch of research has specifically focused on dyadic relationships utilizing regression methods accounting for interdependencies in network data. Here, we find the state-of-the-art *Multiple Regression Quadratic Assignment Procedure* (MRQAP) [10] and its predecessor QAP [5] which permute nodes in such a way that the network structure is kept intact; this allows to test for significance of effects. *Mixed-effects models* [20] add random effects to the models allowing for variation to mitigate non-independence between responses (edges) from the same subject (nodes) [26]. Based on the *quasi essential graph* the work in [14] proposes to compare two graphs (i.e., Bayesian networks) by testing and comparing multiple hypotheses on their edges. Recently, the *generalized hypergeometric ensembles* [2] have been proposed as a framework for model selection and statistical hypothesis testing of finite, directed and weighted networks that allow to encode several topological patterns such as block models where homophily plays an important role in linkage decision. In contrast to our work, neither of these approaches is based on Bayesian hypothesis testing, which avoids some fundamental issues of classic frequentist statistics.

6 Discussion

Next, we discuss some aspects and open questions related to the proposed approach.

Inconsistency of local model. For directed networks, the local ego-network models can assemble a full graph model by defining a probability distribution for the degrees of the source nodes of edges. For undirected networks, this is not directly possible as e.g., the ego-network model for v_A generated an edge from v_A to v_B, but the ego-network model for node v_B did not generate any edge to v_A. Note that this does not affect our comparison of hypotheses as we characterize the network.

Sparse data-connections. Most real networks exhibit small world properties such as high clustering coefficient and fat-tailed degree distributions meaning that the adjacency matrices are sparse. While comparison still relatively judges the plausibility, our hypotheses do not approximate the data curve as shown in Fig. 6b. As an alternative, one might want to limit our beliefs to only those edges that exist in the network, i.e., we would then only build hypotheses on how edge multiplicity varies between edges. Ultimately, our models also warrant extensions to adhere to the degree sequence in the network, e.g., in the direction of multivariate hypergeometric distributions as recently proposed in [2].

Other limitations and future work. The main intent of this work is the introduction of a hypotheses-driven Bayesian approach for understanding edge formation in networks. To that end, we showcased this approach on simple categorical models that warrant extensions, e.g., by incorporating appropriate models for other types of networks such as weighted or temporal networks. We can further investigate how to build good hypotheses by leveraging all node attributes, and infer subnetworks that fit best each of the given hypotheses. Moreover, there can be alternatives for non-attributed networks. For instance, one could use other networks (same nodes, different connections) to verify whether edges from a specific network can be explained by the mechanisms of other networks. In the future, we also plan an extensive comparison to other methods such as MRQAP, mixed-effects models and p^* models.

7 Conclusions

In this paper, we have presented a Bayesian framework that facilitates the understanding of edge formation in attributed multigraphs. The main idea is based on expressing hypotheses as beliefs in parameters (i.e., multiplicity of edges), incorporate them as priors, and utilize Bayes factors for comparing their plausibility. We proposed simple local and global Dirichlet-categorical models and showcased their utility on synthetic and empirical data. For illustration purposes our examples are based on small networks. We tested our approach with larger networks obtaining identical results. In future, our concepts can be extended to further models such as models adhering to fixed degree sequences. We hope that our work contributes new ideas to the research line of understanding edge formation in complex networks.

Acknowledgements This work was partially funded by DFG German Science Fund research projects "KonSKOE" and "PoSTs II".

References

[1] Adamic, L.A., Glance, N.: The political blogosphere and the 2004 us election: divided they blog. In: Proceedings of the 3rd int. workshop on Link discovery, pp. 36–43. ACM (2005)

[2] Casiraghi, G., Nanumyan, V., Scholtes, I., Schweitzer, F.: Generalized hypergeometric ensembles: Statistical hypothesis testing in complex networks. arXiv:1607.02441 (2016)

[3] Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M.: A survey of statistical network models. Foundations and Trends® in Machine Learning **2**(2), 129–233 (2010)

[4] Holland, P.W., Leinhardt, S.: An exponential family of probability distributions for directed graphs. Journal of the american Statistical association **76**(373), 33–50 (1981)

[5] Hubert, L., Schultz, J.: Quadratic assignment as a general data analysis strategy. British journal of mathematical and statistical psychology **29**(2), 190–241 (1976)

[6] Karrer, B., Newman, M.E.: Stochastic blockmodels and community structure in networks. Physical Review E **83**(1), 016,107 (2011)

[7] Kass, R.E., Raftery, A.E.: Bayes factors. Journal of the American Statistical Association **90**(430), 773–795 (1995)

[8] Kim, M., Leskovec, J.: Modeling social networks with node attributes using the multiplicative attribute graph model. In: UAI 2011, Barcelona, Spain, July 14-17, 2011, pp. 400–409 (2011)

[9] Kiti, M.C., Tizzoni, M., Kinyanjui, T.M., Koech, D.C., Munywoki, P.K., Meriac, M., Cappa, L., Panisson, A., Barrat, A., Cattuto, C., et al.: Quantifying social contacts in a household setting of rural kenya using wearable proximity sensors. EPJ Data Science **5**(1), 1 (2016)

[10] Krackhardt, D.: Predicting with networks: Nonparametric multiple regression analysis of dyadic data. Social networks **10**(4), 359–381 (1988)

[11] Kruschke, J.: Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press (2014)

[12] Martin, T., Ball, B., Karrer, B., Newman, M.: Coauthorship and citation patterns in the physical review. Physical Review E **88**(1), 012,814 (2013)

[13] Moreno, S., Neville, J.: Network hypothesis testing using mixed kronecker product graph models. In: Data Mining (ICDM), pp. 1163–1168. IEEE (2013)

[14] Nguyen, H.T.: Multiple hypothesis testing on edges of graph: a case study of bayesian networks

[15] Papadopoulos, F., Kitsak, M., Serrano, M.Á., Boguná, M., Krioukov, D.: Popularity versus similarity in growing networks. Nature **489**(7417), 537–540 (2012)

[16] Pfeiffer III, J.J., Moreno, S., La Fond, T., Neville, J., Gallagher, B.: Attributed graph models: Modeling network structure with correlated attributes. In: WWW, pp. 831–842. ACM (2014)

[17] Robins, G., Pattison, P., Kalish, Y., Lusher, D.: An introduction to exponential random graph (p*) models for social networks. Social networks **29**(2), 173–191 (2007)

[18] Sampson, S.F.: A novitiate in a period of change: An experimental and case study of social relationships. Cornell University (1968)

[19] Schwiebert, L., Gupta, S.K., Weinmann, J.: Research challenges in wireless networks of biomedical sensors. In: Proceedings of the 7th annual international conference on Mobile computing and networking, pp. 151–165. ACM (2001)

[20] Shah, K.R., Sinha, B.K.: Mixed Effects Models, pp. 85–96. Springer New York (1989)

[21] Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. WWW, pp. 1003–1013. ACM (2015)

[22] Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Detecting memory and structure in human navigation patterns using markov chain models of varying order. PloS one **9**(7), e102,070 (2014)

[23] Snijders, T., Spreen, M., Zwaagstra, R.: The use of multilevel modeling for analysing personal networks: Networks of cocaine users in an urban area. Journal of quantitative anthropology **5**(2), 85–105 (1995)

[24] Snijders, T.A.: Statistical models for social networks. Review of Sociology **37**, 131–153 (2011)

[25] Tu, S.: The dirichlet-multinomial and dirichlet-categorical models for bayesian inference. Computer Science Division, UC Berkeley (2014)

[26] Winter, B.: Linear models and linear mixed effects models in r with linguistic applications. arXiv:1308.5499 (2013)

[27] Xiang, R., Neville, J., Rogati, M.: Modeling relationship strength in online social networks. In: WWW, pp. 981–990. ACM (2010)