Ordered community detection in directed networks

Tiago P. Peixoto
Department of Network and Data Science, Central European University, 1100 Vienna, Austria

We develop a method to infer community structure in directed networks where the groups are ordered in a latent one-dimensional hierarchy that determines the preferred edge direction. Our nonparametric Bayesian approach is based on a modification of the stochastic block model (SBM), which can take advantage of rank alignment and coherence to produce parsimonious descriptions of networks that combine ordered hierarchies with arbitrary mixing patterns between groups. Since our model also includes directed degree correction, we can use it to distinguish non-local hierarchical structure from local in- and out-degree imbalance—thus removing a source of conflation present in most ranking methods. We also demonstrate how we can reliably compare with the results obtained with the unordered SBM variant to determine whether a hierarchical ordering is statistically warranted in the first place. We illustrate the application of our method on a wide variety of empirical networks across several domains.

I. INTRODUCTION

Interacting entities in a variety of networked systems form pairwise relationships that are not necessarily symmetric, i.e. an interaction from $i$ to $j$ is distinct from one from $j$ to $i$. Typical examples are predator-prey relationships in food webs [1], antagonist animal behavior [2], reported friendships in social networks [3], and the synaptic connection between neurons. In many such systems, it is often posited that the preferred direction of interaction can be ascribed to an unobserved ordering of the elements involved—placing them on a strict one-dimensional latent hierarchy that most relationships tend to respect. Prominent examples of such ordered systems are species taxa in food webs [4], and dominance hierarchies in animal societies [2].

However, even when present, directed hierarchies are rarely the only dimension that determines how interactions take place. For example, regardless of direction, connections can occur preferentially between specific types of entities, resulting in compartmentalization and heterogeneous mixing patterns that are independent of any underlying ordering. Furthermore, it is also possible for the directed structure of a network not to be associated with any latent hierarchy at all, and to be due instead to entirely different mechanisms. Although in such situations it may still be possible to order the nodes in such a way that the majority of interactions end up respecting a seeming hierarchy, this does not necessarily mean that this is in fact a plausible explanation for how the directions were chosen.

In this work we present a method to infer the ordered modular structure of networks in a manner that simultaneously captures arbitrary mixing patterns and directed hierarchies. Our method is based on a modification of the directed version of the stochastic block model (SBM) [5, 6]—a generative model that can capture arbitrary preferences between groups of nodes. In our modification, the groups themselves are ordered, such that the preferred direction of interaction tends to obey their ranking, while still allowing for the groups to be connected in arbitrary ways, independent of direction. One important ingredient of our model is directed degree-correction [7], which allows nodes that belong to the same group/rank to possess an arbitrarily varied number of incoming and outgoing connections. This means that our method is capable of distinguishing between merely local asymmetries—that stem solely from a node’s tendency to have a particular balance of in and out-connections—and actual hierarchies that affect the structure of the network at a larger scale.

In our methodology we exploit the formal equivalence between statistical inference and data compression [8, 9]. In this setting, we seek to obtain the model inference with the optimal balance between quality of fit and model complexity, such that the amount of information required to describe the network is minimized. This amounts to a nonparametric Bayesian method that can not only determine in a principled manner the most appropriate number of ordered groups, but it also allows us to decide whether a hierarchical structure is warranted at all in the first place, or if we have more evidence instead for a model alternative without any particular ordering between the nodes, but which happens to be more compressive.

Our approach can be compared to previous work in the literature in some important ways. There are several methods that extract relative rankings between the nodes of a network, based on spectral node centrality [10–13], minimum violation ranking [14,15], random utility models [16,21], and latent space models [22,24]. The most central difference between these methods and the one presented in this work is that none of them attempt to simultaneously detect community structure, or include degree-correction. Furthermore, with the exception of the latent space models, these approaches do not attempt to model the placement of the edges, only their latent ordering. Additionally, since they do not attempt to make a statement about data generative processes, they cannot articulate the notion of statistical significance or parsi-
mony [26].

The works that are perhaps closest to ours are the approaches from Letizia et al. [27] and Iacovissi et al. [28]. Letizia et al. [27] considered a ranked SBM with uniform connection probabilities between groups depending only on whether the edge direction violates or not the hierarchy. Besides being unable to uncover heterogeneous mixing patterns and lacking degree correction, the approach of Ref. [27] is not based on a model likelihood, and hence cannot be used to evaluate statistical evidence. The method of Iacovissi et al. [28] is based on a different idea, and combines the SBM with Springrank [24], such that a node can either have a group membership or a ranking, but not both simultaneously. Their model not only lacks degree correction, but its inference is performed in a parametric fashion: the number of groups in the SBM needs to be set a priori, and cannot be extracted from the data itself. Furthermore, the inference procedure developed in Ref. [28] is based on a variational approximation, whereas our approach is based on MCMC using an exact likelihood.

This work is organized as follows. In Sec. II we describe the model and its inference, and in Sec. III we demonstrate how it can be used to simultaneously uncover connection preference and ranking. In Sec. IV we investigate the role of degree-correction in distinguishing local from global ordering, and in Sec. V we consider the problem of model selection between alternatives without latent ordering. We finalize in Sec. VI with a conclusion.

II. NETWORK COMPRESSION VIA MODULAR STRUCTURE, RANK COHERENCE AND ALIGNMENT

We begin by reviewing how the arbitrary mixing pattern between groups of nodes of a directed network can be modelled by the microcanonical degree-corrected stochastic block model (DC-SBM) [29]. In this model, the N nodes are divided into B groups, according to a labelled partition $b = \{ b_i \}$, where $b_i \in [0, B - 1]$ is the group membership of node $i$. As an additional set of parameters, we have the group affinity matrix $e = \{ e_{rs} \}$, where $e_{rs}$ is the number of directed edges that are allowed to exist from group $s$ to $r$, as well as the out- and in-degree sequence $k = \{ (k^{\text{out}}_i, k^{\text{in}}_i) \}$, where $k^{\text{out}}_i$ and $k^{\text{in}}_i$ are the out- and in-degrees of node $i$, respectively. With these constraints in place, a directed multigraph $A = \{ A_{ij} \}$, where $A_{ij}$ is the number of edges from $j$ to $i$, is generated by placing $k^{\text{out}}_i$ and $k^{\text{in}}_i$ “half-edges” on each node $i$, and then pairing them uniformly at random while respecting the counts $e_{rs}$ between all groups $r$ and $s$. A resulting multigraph $A$ is sampled in this manner with probability [29]

$$P(A|k, e, b) = \prod_{rs} e_{rs}! \prod_{ij} A_{ij}! \prod_{r} e^{\text{out}}_r! e^{\text{in}}_r! \prod_{r} k^{\text{in}}_r! k^{\text{out}}_r!$$

(1)

with $e^{\text{out}}_r = \sum_s e_{rs}$ and $e^{\text{in}}_r = \sum_s e_{sr}$, as long as the imposed constraints are respected, otherwise the probability is zero.

The task of identifying the most plausible division of a directed network $A$ into groups consists in inverting the above procedure, and obtaining the posterior distribution

$$P(b|A) = \frac{P(A|b) P(b)}{P(A)}$$

(2)

where $P(b)$ is the prior for the node partition, and $P(A|b)$ is the marginal likelihood,

$$P(A|b) = \sum_{k, e} P(A|k, e, b) P(k, e|b)$$

(3)

$$= P(A|\hat{k}, \hat{e}, b) P(\hat{k}, \hat{e}|b),$$

(4)

where $\hat{k}$ and $\hat{e}$ are the only parameter values compatible with the network $A$ and partition $b$. The prior $P(k, e, b)$ is derived in Ref. [29] and described in Appendix A for completeness. Finding the partition $b$ that maximizes Eq. 2 is equivalent to minimizing the description length of the model $\Sigma$, given by

$$\Sigma(A, b) = - \log_2 P(A|\hat{k}, \hat{e}, b) - \log_2 P(\hat{k}, \hat{e}, b).$$

(5)

The first term in the right hand side of above equation determines the minimum length of a binary message that is required to transmit the matrix $A$, in such a manner that it can be decoded from the message without errors, provided the parameter values $\hat{k}$, $\hat{e}$ and $b$ are already known by the receiver. Likewise, the second term determines the amount of information needed to transmit the model parameters themselves. Therefore, the resulting value $\Sigma(A, b)$ corresponds to the total length of the shortest message that is required to transmit the network $A$ to a receiver that has no prior information on its structure, which must involve sending the parameter values as well.

Minimizing the description length $\Sigma(A, b)$ has the desirable effect of preventing overfitting, which happens for example when we choose a number of groups $B$ that is too large, and the inferred modular structure captures spurious random fluctuations [30]. This is because if a portion of the network (or its entirety) has been generated by a maximally random placement of the edges, it becomes asymptotically impossible to compress it with any algorithm—maximally random data are inherently incompressible [31]. Therefore, if splitting a set of nodes into two groups significantly reduces the description length, this means that the placement of the

---

1 It is possible derive our approach in an entirely equivalent manner by replacing Eq. 1 with independent Poisson distributions for each entry $A_{ij}$, and marginalizing over their parameters [29], but the microcanonical formulation is more convenient for our purposes.
edges involved is very unlikely to have been maximally random, and hence the division is capturing statistically significant structure.

More operationally, the second term in right hand side of Eq. 5 serves as a penalty to the first term, since it tends to increase together with the model complexity, while the first term tends to decrease as the larger number of constraints match the data more closely. The optimal inference is therefore a balance between these two aspects—model complexity and quality of fit—and the overall method serves as formal implementation of Occam’s razor (or the principle of parsimony), which states that simpler models are preferable to more complex ones, provided they have the same explanatory power.

With the posterior of Eq. 2 in place, we can proceed in two ways, depending on our objective. We can find the single partition \( b \) that maximizes that probability, which also minimizes the description length. Alternatively, we can sample partitions from this distribution, and in this way explore the entire landscape of hypotheses, weighted according to their plausibility. The latter can also be seen as a minimum description length (MDL) scheme, with a “one part” description length given by the full marginal distribution, i.e. \( \Sigma(A) = -\log_2 P(A) \), with \( P(A) = \sum_b P(A|b)P(b) \), and noting that \( \Sigma(A) \leq \Sigma(A,b) \), i.e. the full marginal description length is strictly shorter or equal to the one conditioned on a single partition. Both tasks can be accomplished efficiently using Markov chain Monte-Carlo (MCMC), as described in Refs. [32, 33].

A. Identifying group orderings

Although the above model is capable of uncovering directed preferences between groups of nodes, including those where an underlying ordering is present, the ordering itself is not revealed by the model parameters. This is because the posterior distribution of Eq. 2—and therefore also the description length of Eq. 5—is invariant to permutations of the group labels. More specifically, if we consider two partitions \( b \) and \( c \), such that

\[
b_i = \mu(c_i),
\]

where \( \mu(r) \) is a bijection of the group labels, then we have

\[
P(b|A) = P(c|A), \quad \Sigma(A, b) = \Sigma(A, c).
\]

Therefore, the ordering of the groups is entirely immaterial and cannot be used to attain compression under this model, and reveal any aspect of the network structure.

Here we modify precisely this property of the model via a relatively simple, but consequential change. In fact, we keep the model of Eq. 1 exactly as it is, together with the priors for \( k \) and \( b \), and we change only the prior for the group affinities, \( e \). First, we introduce the auxiliary parameter \( m_{rs} \), which counts the total number of edges between groups \( r \) and \( s \) (or twice that number if \( r = s \), regardless of edge direction, i.e.

\[
m_{rs} = \sum_{ij} (A_{ij} + A_{ji}) \delta_{bi,r} \delta_{bj,s}.
\]

Conditioned on this number, we sample the upstream \( (e_{rs}, \text{ with } r > s) \) and downstream \( (e_{rs}, \text{ with } r < s) \) affinities according to

\[
P(e_{rs}, e_{sr}|m_{rs}, p) = \begin{cases} 
\delta_{e_{rs}, m_{rs} - e_{rs}} P(e_{rs}|m_{rs}, p) & \text{if } r < s, \\
\delta_{e_{sr}, m_{rs} - e_{rs}} P(e_{sr}|m_{rs}, p) & \text{if } r > s,
\end{cases}
\]

ensuring that \( e_{rs} + e_{sr} = m_{rs} \), and with the downstream affinity sampled according to a binomial distribution with parameter \( p \),

\[
P(e_{rs}|m_{rs}, p) = \binom{m_{rs}}{e_{rs}} p^{e_{rs}} (1 - p)^{m_{rs} - e_{rs}}.
\]

We call edges that connect nodes of the same group as “lateral,” since they go neither upstream nor downstream. The lateral affinities are given directly by \( m \),

\[
P(e_{rr}|m_{rr}) = \delta_{e_{rr}, m_{rr}/2}.
\]

Introducing the total number of upstream, downstream, and lateral edges,

\[
E^+ = \sum_{r<s} e_{rs}, \quad E^- = \sum_{r<s} e_{sr}, \quad E^0 = \sum_r e_{rr},
\]

respectively, allows us to write the total conditional probability,

\[
P(e|m, p) = \prod_{r<s} \binom{m_{rs}}{e_{rs}} p^{e_{rs}} (1 - p)^{m_{rs} - e_{rs}}.
\]

The parameter \( p \) is considered to be unknown a priori, so we compute the marginal probability,

\[
P(e|m) = \int_0^1 P(e|m, p) P(p) \, dp
\]

\[
= \prod_{r<s} \binom{m_{rs}}{e_{rs}} \left( \frac{E^+ + E^-}{E^+} \right)^{e_{rs}} \cdot \frac{1}{E^+ + E^- + 1},
\]

where we have used a uniform prior density \( P(p) = 1 \). For the symmetric matrix \( m \), we use a uniform distribution conditioned on the total number of edges \( E = E^+ + E^- + E^0 \), given by

\[
P(m|E, B) = \left( \binom{B}{E} \right)^{-1},
\]

where \( \binom{n}{m} = \binom{n + m - 1}{m} \) is the number of \( m \)-combinations from a set of size \( n \), allowing for repetitions. Putting all
rank coherence, as illustrated in Fig. 1. There are two different properties that can make this the description length given by

\[ \Delta_{rs} = e_{rs} - e_{sr}. \]  

(19)

The overall rank alignment is then simply,

\[ \Delta = \sum_{r > s} \Delta_{rs} = E^+ - E^-. \]  

(20)

The larger the magnitude of the overall alignment \( \Delta \), the shortest will be the description length. We can see this by writing the contribution to the description length as

\[ \Sigma(e) = - \sum_{r > s} \log_2 \left( \frac{m_{rs}}{m_{rs}} + \log_2 \left( \frac{E - E^0}{E - E^0 + \Delta} \right) \right) + \log_2(E - E^0 + 1) + \log_2 \left( \frac{B}{2} \right) E, \]  

(21)

where we use the shorthand \( m_{rs} = e_{rs} + e_{sr} \). The maximal rank alignment, \( \Delta = E - E^0 \), achieved with \( \Delta_{rs} = m_{rs} \), will result in the smallest possible description length contribution,

\[ \Sigma(e) = \log_2 \left( \left( \frac{B}{2} \right) E \right) + \log_2(E - E^0 + 1), \]  

(22)

for fixed values of \( B, E \), and \( E^0 \).

Rank coherence, on the other hand, is the uniformity of the values of \( \Delta_{rs} \) across all pairs \((r, s)\). Maximal rank coherence is when all pairwise rank alignments coincide with the overall alignment, i.e.

\[ \Delta_{rs} = \frac{\Delta}{E - E^0} \times m_{rs}, \quad \forall r > s. \]  

(23)

This results in the first term of right hand side of Eq. 21 given by

\[ - \sum_{r < s} \log_2 \left( \frac{m_{rs}}{2(E - E^0)^2 m_{rs}} \right). \]  

(24)

This is the smallest value this term can take, for fixed \( \Delta \) and \( m \) values. Conversely, minimal rank coherence is when the values of \( \Delta_{rs} \) are distributed only between their maximum and minimum values for different \((r, s)\), i.e. \( \Delta_{rs} \in \{m_{rs} - m_{rs}\} \). In this case, the first term will vanish completely from the right hand side of Eq. 21 yielding in a strictly larger description length contribution, if the overall rank alignment \( \Delta \) stays the same. Therefore, rank coherence will always provide improved compression for fixed \( \Delta \) and \( m \) values.

From the above, we can conclude that when rank alignment is maximal, rank coherence must also be maximal, and therefore it amounts for the largest compression possible under this scheme. For intermediary alignment, a
We demonstrate how our model can simultaneously accommodate preference of connections and ranking, by studying the food web of Little Rock Lake [35]. In this network the nodes are taxa, where each taxon is either an individual species, a species subset with distinct set of predators and preys (e.g. different stages of development of individuals of the same species), or an aggregate of similar species. In our representation, a directed edge $i \to j$ exists if taxon $i$ is eaten by taxon $j$. In Fig. 3 we can see the result of our method applied to this network of $N = 183$ nodes. We can identify $B = 22$ ordered taxonomic groups. The vast majority of edges go upstream, revealing a substantial degree of trophic ordering — although the network is far from being acyclic, and we can observe trophic rank violations, cannibalism (self-loops), and lateral predation within the same
trophic group. Overall, the ordering uncovered matches the trophic structure that is well understood for food webs of this type: The basal taxon at the bottom of the hierarchy is an aggregate of microorganisms labelled only “fine organic matter,” which are consumed by a large number of algae species. Intermediary taxa include insects, crustaceans and fish, whereas taxa at the top of the hierarchy correspond to decomposers. However, besides the trophic ordering, we can also identify clear predation hierarchies with opacity indicating probability) and mean value \( \bar{b}_i \) (solid black symbols), for each species.

and a decoupling of rank and group, in the sense that nodes that always belong to different groups can in principle have the same marginal rank distribution. This will happen when the clustering is due predominantly to preference and not a particular position in the hierarchy.

In Fig. 3b we show the marginal rank distribution for the individual taxa, allowing us to identify a fair amount of rank uncertainty at intermediary levels.

### IV. DEGREE CORRECTION: LOCAL VS. GLOBAL ORDERING

We move now to the role of degree correction in our modeling approach. Typical techniques for ordering nodes in a one-dimensional hierarchy attempt, in one way or another, to minimize the rank violations produced by edges that flow in the direction opposite to the rank relationship. As a result, methods of this kind have the tendency to produce orderings that are positively correlated with the difference between out-degree and in-degree of each node,

\[
d_i = k_i^{\text{out}} - k_i^{\text{in}}.
\]

In other words, a node with high out-degree but low in-degree will tend to occupy a low position in hierarchy, whereas a node with low out-degree but high in-degree will tend to occupy a position at the top.

However, we can easily imagine a situation where an arbitrary out-/in-degree sequence leads to an inherent ordering given by \( d_i \), but the edges of the network are
placed otherwise completely at random. In this scenario, this ordering only conveys information about the degree sequence itself, not any additional propensity of placing edges in a manner that respects the ranking of the nodes. Methods that cannot make this distinction will conflate out-/in-degree imbalance with a position in the hierarchy that goes beyond this local property.

Our model allows us to make the distinction between out-/in-degree imbalance and a more meaningful latent hierarchy because it accepts the out-/in-degree sequence \( k \) as a set of parameters that are largely independent from the group affinities \( e \). In this way, it will put nodes in different hierarchical levels only if there is sufficient evidence to justify a preference that goes beyond degree imbalance.

We illustrate this with a simple artificial network model, where all nodes have the same total degree \( k_{\text{out}}^i + k_{\text{in}}^i = k \), but the imbalance is given by an out-degree sampled from a binomial distribution with mean \( (N - i)/(N - 1) \), i.e.

\[
P(k_{\text{out}}^i, k_{\text{in}}^i|k) = \delta_{k_{\text{out}}^i,k-k_{\text{out}}^i} \times \left( \frac{k}{k_{\text{out}}^i} \right) \left( \frac{N - i}{N - 1} \right)^{k_{\text{out}}^i - 1} \left( \frac{i - 1}{N - 1} \right)^{k_{\text{in}}^i - k_{\text{out}}^i}.
\] (29)

Conditioned on a degree sequence sampled in this manner as a hard constraint, we then generate a pairing between the corresponding half-edges uniformly at random, and then obtain a final multigraph \( A \).

When applied to a network sampled from this model, our approach assigns all nodes to a single group—meaning that it (correctly) does not identify any preference of connections that go beyond the degree sequence. As a comparison, we show in Fig. 4a the result obtained with the SpringRank method [24] on the same example. Since this method does not include degree-correction, it also reveals only the degree imbalance. As a means of circumventing the identification of spurious hierarchies of this kind, the authors of Ref. [24] have suggested a null model test, using the rank score provided by the method itself is a test statistic. Unfortunately, this approach is overly sensitive to minor deviations from the null model, as we demonstrate in the following. After generating a network from the above model, we modify the sampled network by adding a small number of random upstream edges involving only the first 5% of the nodes (i.e. nodes with index 1 to \( N/20 \)). The result, as we can see in Fig. 4c, is that the statistical test (correctly) rejects the null model, while the inferred rankings still predominantly reveal only the degree imbalance for

Figure 4. (a) SpringRank values for a network sampled uniformly at random with imposed in/out-degrees themselves sampled from Eq. 29, with \( k = 50 \) and \( N = 1000 \). (b) Same as (a), but with 500 additional upstream edges added uniformly at random between nodes with index in the range \([1, N/20]\) (shown in red). (c) Distribution of SpringRank score values for networks sampled uniformly at random with imposed degree sequence identical to panel (a). The solid vertical line marks the value obtained for the network considered in (b). (d) Marginal rank \( \bar{b}_i \) obtained with the DC-OSBM for the same network as in panel (b).

Figure 5. Comparison of Kendall’s rank correlation coefficient \( \tau \) between the degree imbalance \( d_i \) and rank \( b_i \) for each network in our dataset, for both the degree-corrected and non-degree-corrected version of our model. The sloped dashed line shows the diagonal where the two values are the same.

---

Sampling out-/in-degrees from Eq. 29 may result in values for which the total sum of in- and out-degrees are not identical, which makes a half-edge pairing impossible. If this happens, we resample values for a node chosen uniformly at random, repeatedly, until a feasible degree sequence is obtained.
the majority of the nodes (Fig. 3). This is very much the same problem we encounter when using null model rejection to prevent the detection of spurious communities when doing community detection [26]: the statistical significance of a global quality score tells us very little about the statistical significance of the actual latent variables uncovered—the questions “is the value of the quality score significant?” and “are the inferred latent variables significant?” are not equivalent, and the answer to the first serves as a very poor proxy to the second. Ultimately, the rejection of a null model tells us what kind of structure a network does not have, but cannot tell us what structure it does have. Because of this problem, with a method such as SpringRank, it is not in general possible in uncontrolled empirical settings to fully distinguish between degree imbalance and statistically significant non-local hierarchies.

Since our approach is based on the inference of a flexible generative model, rather than the rejection of a null model, we are able to deal with the above situation in a more satisfying manner. In Fig. 4, we show the inferred rankings of some modified network considered above, according to the degree-corrected ordered SBM (DC-OSBM). Due to degree-correction, the method puts all unperturbed nodes into a single hierarchical level—despite their varied out-/in-degree imbalance—and the perturbed nodes into lower levels, reflecting the upstream edges that were added between them. The interpretation becomes more straightforward: the structure of the first $N/20$ nodes cannot be explained solely by the out-/in-degree imbalance, and the model reveals instead a non-local ordering.

Degree correction is a property that is optional in our approach. It can be “turned off” by choosing an alternative prior for the degree sequence, $P(k|e, b)$ [29]. Therefore, in situations where degree imbalance is expressively desired as a ranking criterion, our method can still be used. However, even with degree-correction, it is still possible to use the degree imbalance to “locally” order nodes that otherwise belong to the same rank, simply by using a lexicographical partial ordering, i.e. $(b_i, d_i) \leq (b_j, d_j)$ if $b_i < b_j$ or $b_i = b_j$ and $d_i \leq d_j$. More importantly, our approach allows for model selection: given the same network $A$, we can decide if the degree-corrected model variant is more compressive or not, by computing its description length, and therefore if there is more statistical evidence justifying its description of the data.

In Fig. 5 we show a comparison between the degree-corrected and non-degree-corrected version of our model for 251 empirical directed networks of different domains (see Appendix C for descriptions). We compute Kendall’s rank correlation coefficient $\tau$ between the degree imbalance $d_i$ and the ranking obtained for each model, for each network in our dataset. The typical case is that the correlation with degree imbalance decreases when degree-correction is used, often substantially, indicating that in those cases the degree sequence is a major contribution to the inferred hierarchy obtained without degree-correction, and there is otherwise no significant support for it. There are also situations when the same correlation values—sometimes also high—are observed for both model variants. This indicates that although the degree sequence itself ends up being informative of the latent hierarchy, this turns out also to be corroborated by an additional alignment with the group ordering that goes significantly beyond the degree imbalance. We can also observe a minority of situations where the correlation increases when degree-correction is employed, but these are mostly due to artefacts caused by the number of hierarchical levels changing significantly from one model to the other.

V. MODEL SELECTION: IS THERE A HIERARCHY?

Given an arbitrary directed network, it is often possible to order its nodes in such a way that the majority of edges ends up following a preferred direction according to that ordering. However, by itself, finding such an ordering is not evidence that it in fact had any role in the formation of the network—in the same manner that finding assortative communities in maximally random networks [30] is not informative of its generative process [26].

A tempting approach to evaluate the statistical significance of a node ordering is to compare it with what can be obtained with a null model, e.g. a network with the same out-/in-degree sequence, but otherwise sampled uniformly at random. This is more easily done via a proxy scalar statistic, such as the total number of rank violations. But as we have already seen in the previous section, this approach, although straightforward, can be quite misleading, since the significance of such global quantities can be very poorly informative of the significance of the actual rankings observed. As seen in Fig. 4, we can obtain overall “significant” results by manipulating only a small minority of the edges of the network. It is important to emphasize that this is not simply a technical problem that can be circumvented by tweaking the test statistic; instead it is a fundamental limitation of null model testing, which is only capable of answering the following question with “yes” or “no”: can the null model be rejected with some confidence? A “no” answer does not give any information about how the null model is likely to be true, and a “yes” answer can tell us nothing more than how the network was not generated—no further details of its generative process can be inferred from this test, including any ranking of its nodes.

A more robust alternative to the rejection of null models is model selection: we articulate a variety of generative models as alternative hypotheses, and check which one is more supported by the data. For the particular problem at hand, we can compare alternative versions of the SBM, containing any combination of degree-correction and latent ordering, in how well they can describe the data. Given the same network $A$ and two
model choices $\mathcal{H}_1$ and $\mathcal{H}_2$, and their uncovered partitions $b^{(1)}$ and $b^{(2)}$, respectively, this comparison is done via the posterior odds ratio,

$$\Lambda = \frac{P(\mathcal{H}_1, b^{(1)} \mid A)}{P(\mathcal{H}_2, b^{(2)} \mid A)} = \frac{P(b^{(1)}, A \mid \mathcal{H}_1)P(\mathcal{H}_1)}{P(b^{(2)}, A \mid \mathcal{H}_2)P(\mathcal{H}_2)}$$

$$= \frac{P(\mathcal{H}_1)}{P(\mathcal{H}_2)} \Sigma_{H_2}(A, b^{(2)}) - \Sigma_{H_1}(A, b^{(1)}),$$

Figure 6. Inferred dominance hierarchy and community structure of antagonistic animal behavior. The columns from left to right contain the results of the non-degree-corrected ordered SBM (OSBM), the degree-corrected ordered SBM (DC-OSBM), and the degree-corrected ordered SBM (DC-SBM). The rows, from top to bottom, show the antagonistic interactions for a group of yellow baboons [38], female bighorn sheep [40], and ant workers [41]. Each panel shows the identified groups for each individual, with the rank labels shown on the nodes—except for the rightmost column, where the groups are not ordered. For the first two leftmost columns, the edge colors indicate the direction: upstream (blue), downstream (red), and lateral (grey). The colors for the rightmost column match the maximum matching with the middle column, and with the unmatched nodes highlighted in red. The panels show also the description length value for each fit.
Figure 7. Comparison between models for 251 empirical directed networks, listed in Appendix C. The values shown are the description length differences with respect to the best model, as indicated in the legend. The networks are ordered by the minimum description length value. The left inset shows the counts that each model type yields a shorter description length, and the right inset shows the distribution of fraction of upstream edges $[E^+/(E^- + E^+)$ or zero if $E = E^-]$, for networks that are best modelled by the DC-OSBM.

with $\Sigma_{H_i}(A, b^{(i)}) = -\log_2 P(b^{(i)}, A|H_i)$ being the description length of the data according to model $H_i$ and its partition $b^{(i)}$. Therefore, if we are a priori agnostic with $P(H_1) = P(H_2)$, we should a posteriori select the model with the shortest description length, and the difference between them will give us the confidence in our selection.

As a case study of the application of the above methodology, we turn to networks of antagonistic behavior between animals [2]. A directed antagonistic relationship between two animals $j \to i$ means that individual $j$ prevails after an aggressive encounter with individual $i$. The overall dominance of $j$ over $i$ is recorded in the multi-graph adjacency matrix $A_{ij}$ as the number of times this particular outcome was observed. Such antagonistic relationships are assumed to reveal a dominance hierarchy in animal societies, the position in which is believed to influence an individual’s access to resources, its chance of survival and reproduction [2].

In Fig. 6, we show the results of some model variants for antagonistic networks of yellow baboons, female bighorn sheep, and ant workers. We consider the non-degree-corrected ordered SBM (OSBM), the degree-corrected ordered SBM (DC-OSBM), and the degree-corrected unordered SBM (DC-SBM). In all cases, the degree-corrected variants yield a shorter description length, indicating that out-/in-degree variability can be largely decoupled from mesoscale mixing patterns. Between the ordered models, the degree-corrected variant yields a smaller number of groups, with a clearer hierarchical structure. However, when compared to the unordered model, the results are mixed. For the yellow baboons, the unordered model yields a significantly improved compression, meaning that heterogeneity of preference and direction of interactions is not optimally captured by the ordered model. This indicates that, although clear asymmetries of outcomes do exist, they cannot be convincingly ascribed to a one-dimensional ordering, even if it simultaneously accounts for group-level preferences. The model variant that discards the inherent ordering can in this case find a more parsimonious description of this network, even though it finds a partition that largely (but not completely) agrees with the ordered model. The results for female bighorn sheep are similar, but far less conclusive: the difference between the description length values from the DC-OSBM and DC-SBM is quite small, yielding only an insignificant posterior odds ratio of $\Lambda \approx 8.6$ in favor of the unordered model. In such a situation we cannot reliably evaluate if the lack of evidence for hierarchy is significant, specially since the partitions yielded by both models differ substantially, and therefore we must conclude that both models offer competing but approximately equally plausible accounts of the data. Finally, the results for the ant worker interactions point in the other direction, and indicate that the ordered model offers a more parsimonious description—indeed in this case the network is completely acyclic, and the inferred model contains only upstream edges.

As the examples above show, the most compressive network representations do not necessarily incorporate rankings between the nodes, although in all cases we can find such an ordering that initially may seem plausible. In Fig. 7 we show a more comprehensive comparison between the ordered and unordered SBMs for a wider set of 251 empirical networks, from diverse domains, listed in Appendix C. For this dataset we find that in fact the DC-OSBM happens to be the most compressive model for a majority of them, with the DC-SBM in the second place. Therefore, it does seem to be the case that node ordering provides opportunities for compression for
many of the networks considered, although the several exceptions mean that ultimately this needs to be evaluated in a case-by-case basis. It is worth observing that even when the ordered model is selected, as we discussed previously, this does not necessarily mean that the rank alignment is large; this could simply be due to an overall rank coherence. Indeed, as we can see in the right inset of Fig. 7, the rank alignment distribution is bimodal, with an abundance of networks with moderate values, and another group with very high values, and hence a more prominent hierarchical structure.

It is useful to remark on the possibly counter-intuitive fact that the ordered versions of the SBM can exploit rank coherence for compression, even when rank alignment is minimal, as we had shown in Fig. 2. This means that in a situation where no actual alignment exists between the group ordering and edge direction, a maximal rank coherence will correspond to a full reciprocity of the edge counts, $e_{rs} = e_{sr}$, which is a special case of the ordered SBMs, but would occur only with a very small probability according to the unordered prior, which expects instead asymmetric matrices. As a result, the ordered SBMs will be selected as the preferred model when a substantial reciprocity between groups exists, which accounts for many cases in Fig. 4. Furthermore, we point out that since the ordered and unordered model versions share the exact same underlying generative model, and differ only in the prior probability for the group affinities, we should not expect any strong general tendency on how many groups are inferred by either variant: If the network has well-defined groups, they will be uncovered by either model. Otherwise, if the groups are not well defined, as is typical for empirical networks that admit different partitions with similar posterior probability [37], the most appropriate model will contribute with a smaller penalty for a subset of them, making them more likely. Whether the selected partitions have more or fewer groups will depend on details of the network structure. We show this in Fig. 5 where it can be seen that the difference in description length between the DC-OSBM and DC-SBM is a relatively poor predictor of which of them uncovers more groups. The larger prevalence of networks for which DC-OSBM simultaneously provides a shorter description length and a larger number of groups when compared to the DC-SBM is better understood as a characteristic of the network corpus considered, rather than a necessary outcome of the comparison between these models.

VI. CONCLUSION

We have demonstrated how a Bayesian version of the directed degree-corrected stochastic block model (DC-SBM)—which is originally invariant to group label permutations—can be suitably modified allowing the relative ordering of the group labels to be used to achieve improved compression whenever the underlying network is embedded in a one-dimensional latent hierarchy, where most edges tend to follow a preferred direction. The resulting ordered SBM can be used to infer latent hierarchies together with arbitrary preferences between groups. We have investigated how degree correction allows the decoupling from out-/in-degree imbalance and latent hierarchies, thus removing a source of conflation that exists in most methods that attempt to rank nodes in a network.

Furthermore, via model selection we showed how it can be determined if the ordering is in fact statistically supported, or if a better description can be obtained with an unordered model. This allows us to evaluate if the ordering obtained is just the necessary outcome of constraints we impose during inference, or if they indeed provide a more plausible description of the data.

It is easy to imagine possible extensions of the ideas presented here that can reveal more detailed relationships between ranking and community structure. For example, in our model, lateral edges (i.e. those that do not involve a difference in rank) can only occur between nodes of the same group. A potential modification would be to allow lateral edges between nodes of different groups. Going further, we could even completely decouple group membership from rank, and infer the relationship between these properties from the data rather than assume it a priori — at the expense of a more complicated model and inference procedure. We leave such possibilities for future work.
[1] R. J. Williams and N. D. Martinez, Simple rules yield complex food webs, *Nature* **404**, 180 (2000).  
[2] E. D. Strauss, J. P. Curley, D. Shizuka, and E. A. Henson, The centennial of the pecking order: current state and future prospects for the study of dominance hierarchies, Philosophical Transactions of the Royal Society  
[3] A. Rapoport and W. J. Horvath, A study of a large sociogram, Behav Sci 6, 279 (1961).  
[4] S. Johnson, V. Domínguez-García, L. Donetti, and M. A. Muñoz, Trophic coherence determines food-web stability, Proceedings of the National Academy of Sciences  
[5] P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic blockmodels: First steps, Social Networks  
[6] T. P. Peixoto, Bayesian Stochastic Blockmodeling, in Advances in Network Clustering and Blockmodeling (John Wiley & Sons, Ltd, 2019) pp. 289–332.  
[7] B. Karrer and M. E. J. Newman, Stochastic blockmodels: First steps, Social Networks  
[8] T. Kawamoto and T. Kobayashi, Sequential locality of graphs and its hypothesis testing, arXiv:2111.11267 [physics] (2021) arXiv: 2111.11267.  
[9] T. P. Peixoto, Descriptive vs. inferential community detection: pitfalls, myths and half-truths, arXiv:2112.00183 [physics, stat] (2022) arXiv:2112.00183.  
[10] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, STANFORD INFOLAB, 17 (1999).  
[11] P. Bonacich, Power and Centrality: A Family of Measures, American Journal of Sociology 92, 1170 (1987).  
[12] F. Fogel, A. d’Aspremont, and M. Vojnovic, Serial-Rank: Spectral Ranking using Seriation, in Advances in Neural Information Processing Systems Vol. 27 (Curran Associates, Inc., 2014).  
[13] M. Cucuringu, Sync-Rank: Robust Ranking, Constrained Ranking and Rank Aggregation via Eigenvector and SDP Synchronization, IEEE Transactions on Network Science and Engineering 3, 58 (2016) conference name: IEEE Transactions on Network Science and Engineering.  
[14] I. Ali, W. D. Cook, and M. Kress, On the Minimum Violations Ranking of a Tournament, Management Science 32, 660 (1986). publisher: INFORMS.  
[15] P. Slater, Inconsistencies in a Schedule of Paired Comparisons, Biometrika 48, 303 (1961). publisher: [Oxford University Press, Biometrika Trust].  
[16] M. Gupte, P. Shankar, J. Li, S. Muthukrishnan, and L. Iftode, Finding hierarchy in directed online social networks, in Proceedings of the 20th international conference on World wide web WWW '11 (Association for Computing Machinery, New York, NY, USA, 2011) pp. 557–566.  
[17] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye, Statistical ranking and combinatorial Hodge theory, Mathematical Programming 127, 203 (2011).  
[18] G. T. Cantwell and C. Moore, Belief propagation for permutations, rankings, and partial orders, arXiv preprint arXiv:2110.00513 (2021).  
[19] R. A. Bradley and M. E. Terry, Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons, *Biometrika* 39, 324 (1952) publisher: [Oxford University Press, Biometrika Trust].  
[20] R. D. Luce, On the possible psychophysical laws, Psychological Review 66, 81 (1959) place: US Publisher: American Psychological Association.  
[21] H. A. DAVID, Ranking from unbalanced paired-comparison data, Biometrika 74, 432 (1987).  
[22] R. J. Williams and D. W. Purves, The probabilistic niche model reveals substantial variation in the niche structure of empirical food webs, Ecology 92, 1849 (2011) eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1890/11-0200.1.  
[23] B. Ball and M. Newman, Friendship networks and social status, Network Science 1, 16 (2013)  
[24] C. De Bacco, D. B. Larremore, and C. Moore, A physical model for efficient ranking in networks, Science Advances 4, eaar8260 (2018) publisher: American Association for the Advancement of Science.  
[25] T. Kawamoto and T. Kobayashi, Sequential locality of graphs and its hypothesis testing, arXiv:2111.11267 [physics] (2021) arXiv: 2111.11267.  
[26] T. P. Peixoto, Hierarchical Block Structures and High-Laplacian Spectra, Physical Review E 95, 012317 (2017)  
[27] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks, Physical Review E 70, 025101 (2004)  
[28] T. Kawamoto and T. Kobayashi, Sequential locality of graphs and its hypothesis testing, arXiv:2111.11267 [physics, stat] (2022) arXiv:2112.00183.  
[29] T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model, Physical Review E 95, 012317 (2017)  
[30] T. P. Peixoto, Nonparametric Bayesian inference of the microcanonical stochastic block model, Physical Review E 95, 012317 (2017)  
[31] T. P. Peixoto, Efficient Monte Carlo and greedy heuristics for the inference of stochastic block models, Physical Review E 89, 012804 (2014)  
[32] T. P. Peixoto, Merge-split Markov chain Monte Carlo for community detection, Physical Review E 102, 012305 (2020)  
[33] T. P. Peixoto, Parsimonious Module Inference in Large Networks, Physical Review Letters 110, 148701 (2013)  
[34] T. P. Peixoto, Hierarchical Block Structures and High-Resolution Model Selection in Large Networks, Physical Review X 4, 011047 (2014)  
[35] N. D. Martinez, Artifacts or Attributes? Effects of Resolution on the Little Rock Lake Food Web, Ecological Monographs 61, 367 (1991)  
[36] T. P. Peixoto, Revealing Consensus and Dissensus between Network Partitions, Physical Review X 11, 021003 (2021) publisher: American Physical Society.
W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57, 97 (1970).

T. P. Peixoto, The graph-tool python library, figshare [10.6084/m9.figshare.1164194] (2014), available at https://graph-tool.skewed.de

T. P. Peixoto, The Netzschleuder network catalogue and repository, (2020), accessible at https://netschleuder.org.

M. D. Domenico, M. A. Porter, and A. Arenas, MuxViz: a tool for multilayer analysis and visualization of networks, Journal of Complex Networks 3, 1 (2015).

R. M. Thompson and C. R. Townsend, Impacts on stream food webs of native and exotic forest: an intercontinental comparison, Ecology 84, 145 (2003).

J. Moody, Peer influence groups: identifying dense clusters in large networks, Social Networks 23, 261 (2001).

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

D. Krackhardt, Cognitive social structures, Social Networks 9, 109 (1987).

D. Sade, Sociometrics of macaca mulatta i. linkages and in gowing matrices, Folia Primatologica 18, 196 (1972).

A. M. Güh, Social behavior of the domestic fowl. Transactions of the Kansas Academy of Science (1903-) 71, 379 (1968).

M. W. Schein and M. H. Fohrman, Social dominance relationships in a herd of dairy cattle, The British Journal of Animal Behaviour 3, 45 (1955).

A.-M. Niekamp, L. A. Merckenn, C. J. Hoeben, and N. H. Dukes-Mujieris, A sexual affiliation network of mature male american bison, Zeitschrift für Tierpsychologie 49, 418 (2010).

L. Šubelj and M. Bajec, Software systems through complex networks science, in Proceedings of the First International Workshop on Software Mining - SoftwareMining’12 (ACM Press, 2012).

L. Šubelj and M. Bajec, Community structure of complex software systems: Analysis and applications, Physica A: Statistical Mechanics and its Applications 390, 2968 (2011).

L. Šubelj and M. Bajec, Clustering assortativity, communities and functional modules in real-world networks (2012).

L. Šubelj, S. Žitnik, N. Blagus, and M. Bajec, Node mixing and group structure of complex software networks, Advances in Complex Systems 17, 1450022 (2014).

V. Gelardi, J. Godard, D. Paleressompoulle, N. Claidiere, and A. Barrat, Measuring social networks in primates: wearable sensors versus direct observations, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 476, 20190737 (2020).

J. S. Coleman et al., Introduction to mathematical sociology., Introduction to mathematical sociology. (1964).

L. M. Gerdes, K. Ringler, and B. Autin, Assessing the abu sayyaf group's strategic and learning capacities, Studies in Conflict & Terrorism 37, 267 (2014).

G. G. V. D. Bunt, M. A. V. Duijn, and T. A. Snijders, Computational & Mathematical Organization Theory 5, 167 (1999).

J. Coleman, E. Katz, and H. Menzel, The Diffusion of an Innovation Among Physicians, Sociometry 20, 253 (1957).

S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Network motifs in the transcriptional regulation network of Escherichia coli, Nature genetics 31, 64 (2002).

L. M. Fedigan and P. J. Asquith, The monkeys of Arashiyama: thirty-five years of research in Japan and the West (SUNY Press, 1991).

W. Raub, Emmanuel lazeva: The collegial phenomenon. the social mechanisms of cooperation among peers in a corporate law partnership. oxford: Oxford university press, 2001. xi + 346 pp, European Sociological Review 21, 183 (2005).

R. E. Ulanowicz and D. L. DeAngelis, Network analysis of tropic dynamics in south florida ecosystems, US Geological Survey Program on the South Florida Ecosystem , 114 (1999).

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network Motifs: Simple Building Blocks of Complex Networks, Science 298, 824 (2002).

A. Clauset, S. Arbesman, and D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks, Science Advances 1, 10.1126/sciadv.1400005 (2015).

K. Ryan, Z. Lu, and T. A. Meirnertzhagen, The CNS connectome of a tadpole larva of Ciona intestinalis (L.) highlights sidedness in the brain of a chordate sibling, eLife 5, e16962 (2016).

The structure of the nervous system of the nematode-Caenorhabditis elegans, Philosophical Transactions of the Royal Society of London. B, Biological Sciences 314, 1 (1986).

D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393, 409 (1998).

S. J. Cook, T. A. Jarrell, C. A. Brittin, Y. Wang, A. E. Bioniarz, M. A. Yakovlev, K. C. Q. Nguyen, L. T.-H. Tang, E. A. Bayer, J. S. Duerr, H. E. Bïlow, O. Hobert,
D. H. Hall, and S. W. Emmons, Whole-animal connectomes of both caenorhabditis elegans sexes, Nature 571, 63 (2019).

[76] L. C. Freeman, C. M. Webster, and D. M. Kirke, Exploring social structure using dynamic three-dimensional color images, Social Networks 20, 109 (1998).

[77] J. A. Dunne, C. C. Labandeira, and R. J. Williams, Highly resolved early eocene food webs show development of modern trophic structure after the endcretaceous extinction, Proceedings of the Royal Society B: Biological Sciences 281, 20133280 (2014).

[78] R. Michalski, S. Palus, and P. Kazienko, Matching organizational structure and social network extracted from email communication, in Business Information Systems (Springer Berlin Heidelberg, 2011) pp. 197–206.

[79] R. M. Ewing, P. Chu, F. Elisma, H. Li, P. Taylor, S. Climie, L. McBroom-Cerjawski, M. D. Robinson, L. O’Connor, M. Li, R. Taylor, M. Dharsee, Y. Ho, A. Heilbut, L. Moore, S. Zhang, O. Ornatsky, Y. V. Bukhman, M. Ethier, Y. Sheng, J. Vasilescu, M. Abu-Farha, J.-P. Lambert, H. S. Duewel, I. I. Stewart, B. Kuehl, K. Hogue, K. Colwill, K. Gladwish, B. Muskat, R. Kinach, S.-L. Adams, M. F. Moran, G. B. Morin, T. Topaloglu, and D. Figeys, Large-scale mapping of human protein–protein interactions by mass spectrometry, Molecular Systems Biology 3, 89 (2007).

[80] U. Stefl, U. Worn, M. Lalowski, C. Haenig, S. K. Blackburn, P. Puzis, R. G. Guimerà, L. Danon, A. Diaz-Guilera, F. Giralt, and R. Milo, The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives, in Proceedings of the 9th International Symposium on String Processing and Information Retrieval, SPIRE 2002 (Springer-Verlag, London, UK, UK, 2002) pp. 1–10.

[81] F. Gursoy and D. Gunnecek, Influence maximization in social networks under deterministic linear threshold model, Knowledge-Based Systems 161, 111 (2018).

[82] M. Fire, R. Puzis, and Y. Elovici, Link prediction in highly fractional data sets, in Handbook of Computational Approaches to Counterterrorism (Springer New York, 2012) pp. 283–308.

[83] Kaggle, Chess ratings - elo versus the rest of the world, https://www.kaggle.com/c/chess/data

[84] J. Leskovec, D. Huttenlocher, and J. Kleinberg, Signed networks in social media, in ICWSM 2010 - International AAAI Conference on Weblogs and Social Media (2010).

[85] The caida as relationships dataset, http://www.caida.org/data/as-relationships/

[86] K. Guilkisson, Python dependency analysis, http://kgullikson88.github.io/blog/pypi-analysis.html

[87] G. Palla, I. J. Farkas, P. Pollner, I. Derényi, and T. Vicsek, Directed network modules, New Journal of Physics 9, 186 (2007).

[88] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, Information Retrieval 3, 127 (2000).

[89] T. A. Dallas, A. A. Aguirre, S. Budischak, C. Carlson, V. Ewenza, B. Han, S. Huang, and P. R. Stephens, Gauging support for macroecological patterns in helminth parasites, Global Ecology and Biogeography 27, 1437 (2018).

[90] E. Rochlo, Map of the fediverse, (2018).

[91] O. Richters and T. P. Peixoto, Trust Transitivity in Social Networks, PLoS ONE 6, e18384 (2011).

[92] J. H. Fowler and S. Jeon, The authority of supreme court, Political Analysis 15, 324 (2007).

[93] O. Richters and T. P. Peixoto, Trust Transitivity in Social Networks, PLoS ONE 6, e18384 (2011).

[94] J. Leskovec, D. Huttenlocher, and J. Kleinberg, Signed networks in social media, in ICWSM 2010 - International Approaches to Counterterrorism

[95] J. Gehrke, P. Ginsparg, and J. Kleinberg, Overview of the 2nd ACM Workshop on Online Social Networks in Social Media, in Autonomic and Secure Computing 8th International Conference on Dependable, Autonomic and Secure Computing (2009) pp. 658–663.

[96] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, Network analysis and the law: Measuring the legal importance of precedents at the u.s. supreme court, Political Analysis 15, 324 (2007).

[97] B. Klimt and Y. Yang, The Enron Corpus: A New Dataset for Email Classification Research, in Machine Learning: ECML 2004, Lecture Notes in Computer Science No. 3201, edited by J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi (Springer Berlin Heidelberg, 2004) pp. 217–226.

[98] J. Gehrke, P. Ginsparg, and J. Kleinberg, Overview of the 2003 KDD Cup, SIGKDD Explor. News. 5, 149 (2003).

[99] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data 1, 10.1145/1217299.1217301 (2007).

[100] J. A. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data 1, 10.1145/1217299.1217301 (2007).

[101] M. Ley, The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives, in Proceedings of the 9th International Symposium on String Processing and Information Retrieval, SPIRE 2002 (Springer-Verlag, London, UK, UK, 2002) pp. 1–10.

[102] F. Gursoy and D. Gunnecek, Influence maximization in social networks under deterministic linear threshold model, Knowledge-Based Systems 161, 111 (2018).

[103] M. Fire, R. Puzis, and Y. Elovici, Link prediction in highly fractional data sets, in Handbook of Computational Approaches to Counterterrorism (Springer New York, 2012) pp. 283–308.

[104] J. Leskovec, D. Huttenlocher, and J. Kleinberg, Signed networks in social media, in ICWSM 2010 - International Approaches to Counterterrorism

[105] J. Gehrke, P. Ginsparg, and J. Kleinberg, Overview of the 2nd ACM Workshop on Online Social Networks in Social Media, in Autonomic and Secure Computing 8th International Conference on Dependable, Autonomic and Secure Computing (2009) pp. 658–663.

[106] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, Network analysis and the law: Measuring the legal importance of precedents at the u.s. supreme court, Political Analysis 15, 324 (2007).

[107] B. Klimt and Y. Yang, The Enron Corpus: A New Dataset for Email Classification Research, in Machine Learning: ECML 2004, Lecture Notes in Computer Science No. 3201, edited by J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi (Springer Berlin Heidelberg, 2004) pp. 217–226.

[108] J. Gehrke, P. Ginsparg, and J. Kleinberg, Overview of the 2003 KDD Cup, SIGKDD Explor. News. 5, 149 (2003).

[109] J. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data 1, 10.1145/1217299.1217301 (2007).

[110] J. A. Leskovec, J. Kleinberg, and C. Faloutsos, Graph evolution: Densification and shrinking diameters, ACM Trans. Knowl. Discov. Data 1, 10.1145/1217299.1217301 (2007).

[111] M. Ley, The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives, in Proceedings of the 9th International Symposium on String Processing and Information Retrieval, SPIRE 2002 (Springer-Verlag, London, UK, UK, 2002) pp. 1–10.
Appendix A: The directed, degree-corrected SBM

As shown in the main text, and derived in Ref. [29], the microcanonical directed degree-corrected SBM has a likelihood given by

$$P(A|k, e, b) = \prod_r \varepsilon_{rs}! \prod_i k_{i}^{\text{out}}! k_{i}^{\text{in}}! \prod_{ij} A_{ij}! \prod_{r} e_{r}^{\text{out}}! e_{r}^{\text{in}}!$$  \hspace{1cm} (A1)

where $k = \{(k_{i}^{\text{out}}, k_{i}^{\text{in}})\}$ is the imposed out-/in-degree sequence, with

$$k_{i}^{\text{out}} = \sum_j A_{ji}, \quad k_{i}^{\text{in}} = \sum_j A_{ij}$$  \hspace{1cm} (A2)

and $e = \{e_{rs}\}$ being the edge counts between groups, with marginals given by

$$e_{r}^{\text{out}} = \sum_s e_{rs}, \quad e_{r}^{\text{in}} = \sum_s e_{sr}.$$  \hspace{1cm} (A3)

The prior for out-/in-degree sequence is conditioned on the out-/in-degree distributions $\eta^{\text{out}} = \{\eta_{k}^{\text{out}}\}$ and $\eta^{\text{in}} = \{\eta_{k}^{\text{in}}\}$, where $\eta_{k}^{\text{out}}$ ($\eta_{k}^{\text{in}}$) is the number of nodes in group $r$ with out-degree (in-degree) equal to $k^{\text{out}}$ ($k^{\text{in}}$), and is given by

$$P(k|\eta^{\text{out}}, \eta^{\text{in}}) = \prod_r \left[ \prod_{k^{\text{out}}} \frac{\eta_{k}^{\text{out}}}{n_{r}^{k^{\text{out}}}}! \right] \left[ \prod_{k^{\text{in}}} \frac{\eta_{k}^{\text{in}}}{n_{r}^{k^{\text{in}}}}! \right]$$  \hspace{1cm} (A4)

with $n_{r}$ being the number of nodes in group $r$. The out-/in-degree distributions themselves sampled from group-wise uniform distributions,

$$P(\eta^{\text{out}}, \eta^{\text{in}}|e, b) = \prod_r q(e_{r}^{\text{out}}, n_{r})^{-1} q(e_{r}^{\text{in}}, n_{r})^{-1},$$  \hspace{1cm} (A5)

where $q(m, n) = q(m, n-1) + q(m-n, n)$, with boundary conditions $q(m, 1) = 1$ for $m > 0$ and $q(m, n) = 0$ for $m \leq 0$ or $n \leq 0$, is the number of restricted partitions of the integer $m$ into at most $n$ parts.

The non-degree-corrected version of the model can be obtained by replacing the above prior for $k$ with

$$P(k|e, b) = \prod_r n_{r}^{k^{\text{out}}!} \prod_i (k_{i}^{\text{out}})! b_{r,i}^{k_{i}^{\text{out}}} \times n_{r}^{k^{\text{in}}!} \prod_i (k_{i}^{\text{in}})! b_{r,i}^{k_{i}^{\text{in}}}.$$  \hspace{1cm} (A6)

For the partition we have the prior

$$P(b) = P(b|n)P(n|B)P(B)$$  \hspace{1cm} (A7)

$$= \frac{\prod_r n_{r}!}{N!} \left( \frac{N-1}{B-1} \right)^{-1} \frac{1}{N},$$  \hspace{1cm} (A8)

Finally, for the edge counts we have a uniform distribution

$$P(e|B) = \left( \frac{B!}{E} \right)^{-1},$$  \hspace{1cm} (A9)

or a nested sequence of SBMs for the nested version of the model, as described in Ref. [29].

Appendix B: MCMC inference

The inference procedure we use in this work is Markov chain Monte Carlo (MCMC), implemented as follows. Starting from a partition $b$, a new partition $b'$ is proposed with probability $P(b'|b)$ and accepted according to the Metropolis-Hastings criterion [12, 13], i.e. with a probability given by

$$\min \left( \frac{P(b'|A)P(b'|b)}{P(b|A)P(b|b)} \right) \leq 1,$$  \hspace{1cm} (B1)

otherwise it is rejected. If the move proposals are ergodic and aperiodic, repeating the above procedure will eventually sample partitions from the target distribution $P(b|A)$, which needs to be computed only up to a normalization constant. The move proposals we use are the merge-split moves described in Ref. [33] which have very good mixing properties, and allow each sweep of the algorithm (i.e. a number of moves that allow each node to change its membership at least once) to be computed in linear time $O(N + E)$, independent on the number of groups being considered at any given time.

Although the above method can be used indistinguishably for the ordered and unordered SBMs, it is beneficial to modify it in a subtle way for the ordered variant. Since the unordered SBM is invariant to label permutations, the implementation of the above algorithm can be done without taking into consideration which labels are used when a new group is created. On the other hand, with the ordered model, the relative ordering of the newly created group becomes important. Instead of using the
numeric value of the label itself, it is fact more efficient to associate with each label \( r \) an auxiliary real numeric value \( u_r \in [0, 1] \) which establishes its ordering, i.e. \( r < s \) if and only if \( u_r < u_s \). Thus, whenever a new group \( r \) is created, its relative placement is given a new value \( u_r \) sampled uniformly at random in the interval \([0, 1]\). The ergodicity of this auxiliary variable is preserved by allowing the move of the nodes of a group \( r \) to a newly created group \( s \), with a new value of \( u_s \). In this way, we can sample re-orderings of the group labels without actually having to change them.

The above approach will sample partitions from the posterior distribution. To obtain the partition that maximizes it, we need simply to add an inverse temperature parameter \( \beta \), i.e. \( P(b|A) \rightarrow P(b|A)^{\beta} \), and compute the limit \( \beta \rightarrow \infty \), which means we only accept a move proposal if it strictly increases the posterior probability.

A C++ implementation of the above algorithm is available as part of the graph-tool library [41].

Appendix C: Network data

In table [1] we list the network data used in this work, which are freely available from the Netschleuder repository [45].
Table I: Directed network data used in this work, indexed in increasing order of minimum description length (in accordance with Fig. 7), together with the number of nodes $N$ and edges $E$, the description length in bits obtained with the four model variants, as well as the model with the shortest description length.