Fast and scalable likelihood maximization for Exponential Random Graph Models with local constraints

Nicolò Vallarano, 1 Matteo Bruno, 1 Emiliano Marchese, 1 Giuseppe Trapani, 1 Fabio Saracco, 1 Giulio Cimini, 2 Mario Zanon, 1 and Tiziano Squartini 1, 3

1 IMT School for Advanced Studies Lucca, P.zza San Francesco 19, 55100 Lucca (Italy)
2 Physics Department and INFN, ‘Tor Vergata’ University of Rome, 00133 Rome (Italy)
3 Institute for Advanced Study (IAS), University of Amsterdam, Oude Turfmarkt 145, 1012 GC Amsterdam (The Netherlands)

Abstract. Exponential Random Graph Models (ERGMs) have gained increasing popularity over the years. Rooted into statistical physics, the ERGMs framework has been successfully employed for reconstructing networks, detecting statistically significant patterns in graphs, counting networked configurations with given properties. From a technical point of view, the ERGMs workflow is defined by two subsequent optimization steps: the first one concerns the maximization of Shannon entropy and leads to identify the functional form of the ensemble probability distribution that is maximally non-committal with respect to the missing information; the second one concerns the maximization of the likelihood function induced by this probability distribution and leads to its numerical determination. This second step translates into the resolution of a system of $O(N)$ nonlinear, coupled equations (with $N$ being the total number of nodes of the network under analysis), a problem that is affected by three main issues, i.e. accuracy, speed and scalability. The present paper aims at addressing these problems by comparing the performance of three algorithms (i.e. Newton’s method, a quasi-Newton method and a recently-proposed fixed-point recipe) in solving several ERGMs, defined by binary and weighted constraints in both a directed and an undirected fashion. While Newton’s method performs best for relatively little networks, the fixed-point recipe is to be preferred when large configurations are considered, as it ensures convergence to the solution within seconds for networks with hundreds of thousands of nodes (e.g. the Internet, Bitcoin). We attach to the paper a Python code implementing the three aforementioned algorithms on all the ERGMs considered in the present work.

PACS numbers: 02.10.Ox, 89.75.Hc, 02.70.Rr, 05.10.-a

INTRODUCTION

Over the last 20 years, network theory has emerged as a successful framework to address problems of scientific and societal relevance [1]: examples of processes that are affected by the structural details of the underlying network are provided by the spreading of infectious diseases [2–4], opinion dynamics [5], the propagation of losses during financial crises [6], etc.

Within such a context, two needs have emerged quite naturally [7]: 1) detecting the topological properties of a real network that can be deemed as statistically significant - typically those higher-order properties (e.g. the assortativity, the clustering coefficient, etc.) that may be explained by local features of the nodes such as the degrees, 2) inferring the relevant details of a given network structure in case only partial information is available. Both goals can be achieved by constructing a framework for defining benchmarks, i.e. synthetic configurations retaining only some of the properties of the original system - the so-called constraints - and, otherwise, being maximally random.

Two different kinds of approaches have been proposed so far, i.e. the microcanonical and the canonical ones. Microcanonical approaches [8–14] artificially generate many randomized variants of the observed network by enforcing the constraints in a ‘hard’ fashion, i.e. by creating configurations on each of which the constrained properties are identical to the empirical ones. On the other hand, canonical approaches [15–19] enforce constraints in a ‘soft’ fashion, by creating a set of configurations over which the constrained properties are on average identical to the empirical ones. Softening the requirement of matching the constraints has a clear advantage: allowing the mathematical expression for the probability of a generic configuration, $G$, to be obtained analytically, as a function of the enforced constraints.

In this second case, a pivotal role is played by the formalism of the Exponential Random Graph Models (ERGMs) [20] whose popularity has steadily increased over the years. The ERGMs mathematical framework dates back to Gibbs’ (re)formulation of statistical mechanics and is based upon the variational principle known as maximum entropy, stating that the probability distribution that is maximally non-committal with respect to the missing information is the one maximizing the Shannon entropy [21]. This allows self-consistent inference to be made, by assuming maximal ignorance about the unknown degrees of freedom of the system.

In the context of network theory, the ERGMs frame-
work has found a natural application to the resolution of the two aforementioned problems, i.e. 1) that of defining null models of the original network, in order to assess the significance of empirical patterns - against the hypothesis that the network structure is solely determined by the constraints themselves and 2) that of deriving the most probable configurations that are compatible with the available details about a specific network.

In both cases, after the functional form of the probability distribution has been identified, via the maximum entropy principle, one also needs to numerically determine it: to this aim, the likelihood maximization principle can be invoked, stating to maximize the probability of observing the actual configuration. This prescription typically translates into the resolution of a system of $O(N)$ non-linear, coupled equations - with $N$ representing the number of nodes of the network under analysis.

Problems like these are usually affected by the issues of accuracy, speed and scalability: the present paper aims at addressing them at once, by comparing the performance of three algorithms, i.e. Newton’s method, a quasi-Newton method and a recently-proposed fixed-point recipe [22, 23], to solve a variety of ERGMs, defined by binary and weighted constraints in both a directed and an undirected fashion.

We would like to stress that, while the theoretical architecture of ERGMs is well established, an exhaustive study of the computational cost required for their numerical optimization, alongside with a comparison among the most popular algorithms designed for the task, is still missing. Our work aims at filling precisely this gap. Additionally, we provide a Python code implementing the three aforementioned recipes on all the ERGMs considered in the present work.

GENERAL THEORY

Canonical approaches aim at obtaining the mathematical expression for the probability of a generic configuration, $G$, as a function of the observed constraints: ERGMs realize this by maximizing the Shannon entropy [15, 16].

The Maximum Entropy Principle

Generally speaking, the problem to be solved in order to find the functional form of the probability distribution to be employed as a benchmark reads

$$\arg \max_P S[P] \quad \text{s.t.} \quad \sum_G P(G)C_i(G) = \langle C_i \rangle, \quad i = 0 \ldots M$$

where Shannon entropy reads

$$S[P] = -\sum_G P(G) \ln P(G)$$

and $\bar{C}(G)$ is the vector of constraints representing the information defining the benchmark itself (notice that $C_0 = \langle C_0 \rangle = 1$ sums up the normalization condition). The solution to the problem above can be found by maximizing the Lagrangian function

$$\mathcal{L}(P, \theta) \equiv S[P] + \sum_{i=0}^M \theta_i \left[ -\sum_G P(G)C_i(G) + \langle C_i \rangle \right]$$

with respect to $P(G)$. As a result one obtains

$$P(G|\bar{\theta}) = \frac{e^{-\mathcal{H}(G, \bar{\theta})}}{Z(\bar{\theta})}$$

with $\mathcal{H}(G, \bar{\theta}) = \theta \cdot \bar{C}(G) = \sum_{i=1}^M \theta_i C_i(G)$ representing the Hamiltonian, i.e. the function summing up the proper, imposed constraints and $Z(\bar{\theta}) = \sum_G P(G) = \sum_G e^{-\mathcal{H}(G, \bar{\theta})}$ representing the partition function, ensuring that $P(G)$ is properly normalized. Constraints play a pivotal role, either representing the information to filter, in order to assess the significance of certain quantities, or the only available one, in order to reconstruct the inaccessible details of a given configuration.

The Maximum Likelihood Principle

The formalism above is perfectly general; however, it can be instantiated to study an empirical network configuration, say $G^*$. In this case, the Lagrange multipliers ‘acting’ as unknown parameters in eq. (4) can be numerically estimated by maximizing the associated likelihood function [15, 24]. The latter is defined as

$$\mathcal{L}(\bar{\theta}) \equiv \ln P(G^*|\bar{\theta}) = -\mathcal{H}(G^*, \bar{\theta}) - \ln Z(\bar{\theta})$$

and must be maximized with respect to the vector $\bar{\theta}$. Remarkably, whenever the probability distribution is exponential (as the one deriving from the Shannon entropy maximization), the likelihood maximization problem

$$\arg \max_{\bar{\theta}} \mathcal{L}(\bar{\theta})$$

is characterized by first-order necessary conditions for optimality reading
Combining the MEP and the MLP principles

The Maximum Entropy Principle (MEP) and the Maximum Likelihood Principle (MLP) encode two different prescriptions aiming, respectively, at determining the functional form of a probability distribution and its numerical value. In optimization theory, the problem \([1]\) is known as *primal problem*: upon noticing that the Shannon entropy is concave, while the imposed constraints are linear in \(P(G)\), one concludes that the primal problem is convex (it is easy to see this, by rewriting it as a minimization problem for \(-S[P]\)).

As convexity implies *strong duality*, we can, equivalently, consider an alternative version of the problem to optimize, known as *dual problem*. In order to define it, let us consider the Lagrangian function

\[
\mathcal{L}(P, \mathbf{\theta}) = S[P] + M \sum_{i=1}^{M} \theta_i \left[ - \sum_{G} P(G) C_i(G) + C_i(G^*) \right] \tag{10}
\]

where, now, the generic expectation of the \(i\)-th constraint, \(\langle C_i \rangle\), has been replaced by the corresponding empirical value, \(C_i(G^*)\). As the dual function is given by

\[
P(G^*|\mathbf{\theta}) \equiv \arg \max_P \mathcal{L}(P, \mathbf{\theta}), \tag{11}
\]

the dual problem reads

\[
\nabla_{\mathbf{\theta}} L(\mathbf{\theta}) = \partial \mathcal{L}(\mathbf{\theta}) / \partial \theta_i = -C_i(G^*) - \sum_{G} C_i(G) P(G) = -C_i(G^*) + \langle C_i \rangle = 0, \quad i = 1 \ldots M \tag{7}
\]

and leading to the system of equations

\[
\nabla \mathcal{L}(\mathbf{\theta}) = \mathbf{0} \implies \mathbf{\bar{C}}(G^*) = \langle \mathbf{\bar{C}} \rangle \tag{8}
\]

to be solved. These conditions, however, are sufficient to characterize a maximum only if \(\mathcal{L}(\mathbf{\theta})\) is concave. This is indeed the case, as we prove by noticing that

\[
H_{ij} = \nabla^2_{\theta_i \theta_j} \mathcal{L}(\mathbf{\theta}) = \partial^2 \mathcal{L}(\mathbf{\theta}) / \partial \theta_i \partial \theta_j = -\partial^2 \ln Z(\mathbf{\bar{\theta}}) / \partial \theta_i \partial \theta_j = -\text{Cov}[C_i, C_j], \quad i, j = 1 \ldots M \tag{9}
\]

i.e. that the Hessian matrix, \(\mathbf{H}\), of our likelihood function is ‘minus’ the covariance matrix of the constraints, hence negative semidefinite by definition. The fourth passage is an example of the well-known fluctuation-response relation \([20]\).

A graphical representation of how the two principles work is shown in fig. 1.
FIG. 2. System diagram illustrating the models discussed in the present paper and implemented in the ‘NEMTROPY’ package, attached to it: it represents a sort of guide to individuate the best model for analysing the system at hand. Our package handles both monopartite and bipartite networks; while the latter ones have been considered only in their binary, undirected fashion, the former ones can be modeled either in a binary or a weighted fashion, allowing for both undirected and directed links.

\[
\text{arg max} \quad \text{arg min} \quad -\mathcal{L}(P(\tilde{\theta}), \tilde{\theta}) \quad (12)
\]

which is a convex problem by construction; this is readily seen by substituting eq. (4) into eq. (10), operation that leads to the expression

\[
-\mathcal{L}(P(\tilde{\theta}), \tilde{\theta}) = -\tilde{\theta} \cdot \tilde{C}(G^*) - \ln Z(\tilde{\theta}) = \mathcal{L}(\tilde{\theta}), \quad (13)
\]
i.e. the likelihood function introduced in eq. (6). In other words, eq. (12) combines the MEP and the MLP into a unique optimization step whose score function becomes the Lagrangian function defined in eq. (10).

Optimization algorithms for non-linear problems

In general, the optimization problem defined in eq. (12) cannot be solved analytically, whence the need to resort to numerical methods. For an exhaustive review on numerical methods for optimization we refer the interested reader to [25, 26]: in the following, we present only the concepts that are of some relevance for us. The problem

\[
\text{arg max} \quad \mathcal{L}(\tilde{\theta}) \quad (14)
\]
is a Nonlinear Programming Problem (NLP). In order to solve it numerically, we will adopt a Sequential Quadratic Programming (SQP) approach. Starting from an initial guess \(\tilde{\theta}^{(0)}\), SQP solves eq. (14) by iteratively updating the vector of Lagrange multipliers

\[
\tilde{\theta}_i^{(n+1)} = \tilde{\theta}_i^{(n)} + \alpha \Delta \tilde{\theta}_i^{(n)}, \quad i = 1 \ldots M \quad (15)
\]
according to the rule

\[
\Delta \tilde{\theta}_i^{(n)} = \arg \max_{\Delta \tilde{\theta}_i} \left[ \nabla_{\tilde{\theta}_i} \mathcal{L}(\tilde{\theta}) \Delta \tilde{\theta}_i + \frac{1}{2} \sum_{j,k} \Delta \tilde{\theta}_j H_{jk}^{(n)} \Delta \tilde{\theta}_k \right] \quad (16)
\]
\forall i, leading to the set of equations

\[
\nabla_{\tilde{\theta}_i} \mathcal{L}(\tilde{\theta}) + \sum_{j} H_{ij}^{(n)} \Delta \tilde{\theta} = 0, \quad i = 1 \ldots M \quad (17)
\]
which can be compactly rewritten as
Table 1. Graphical visualization of the constraints defining the ERGMs considered in this work. Notice that while the enhanced models (i.e. UECM and DECM) constrain binary and weighted quantities in a joint fashion, the conditional models (i.e. the CReM ones) allow for a ‘separate’ specification of them.

| Model   | Constraints                  | Visualization |
|---------|------------------------------|---------------|
| UBCM    | Degree Sequence              | [              ] |
| DBCM    | Out-Degree Sequence          | [              ] |
|         | In-Degree Sequence           | [              ] |
| BiCM    | Layer 1 Degree Sequence      | [              ] |
|         | Layer 2 Degree Sequence      | [              ] |
| UECM    | Degree Sequence              | [              ] |
|         | Strength Sequence            | [              ] |
| DECM    | Out-Degree Sequence          | [              ] |
|         | In-Degree Sequence           | [              ] |
|         | Out-Degree Sequence          | [              ] |
|         | In-Degree Sequence           | [              ] |
| CReM    | Strength Sequence            | [              ] |
| Undirected | conditional to               | [              ] |
|         | Degree Sequence              | [              ] |
| CReM    | Out-Strength Sequence        | [              ] |
| Directed | conditional to               | [              ] |
|         | Out-Degree Sequence          | [              ] |
|         | In-Degree Sequence           | [              ] |

FIG. 3. Graphical visualization of the constraints defining the ERGMs considered in this work. Notice that while the enhanced models (i.e. UECM and DECM) constrain binary and weighted quantities in a joint fashion, the conditional models (i.e. the CReM ones) allow for a ‘separate’ specification of them.

\[
\Delta \theta^{(n)} = -H^{(n)}^{-1}\nabla \mathcal{L}(\theta). \tag{18}
\]

The stepsize \( \alpha \in (0, 1) \) is selected to ensure that \( \mathcal{L}(\hat{\theta}^{(n+1)}) > \mathcal{L}(\hat{\theta}^{(n)}) \) via a backtracking, line search procedure: starting from \( \alpha = 1 \), if the Armijo condition

\[
\mathcal{L}(\hat{\theta}^{(n)}) + \alpha \Delta \hat{\theta}^{(n)} < \mathcal{L}(\hat{\theta}^{(n)}) + \gamma \alpha \nabla \mathcal{L}(\hat{\theta})^\top \Delta \hat{\theta}, \tag{19}
\]

is violated, we set \( \alpha \leftarrow \beta \alpha \) (\( \gamma \in (0, 0.5) \) and \( \beta \in (0, 1) \) are the parameters of the algorithm). On the other hand, the term \( H^{(n)} \) can be selected according to a variety of methods. In the present contribution we focus on the
following three ones.

**Newton’s method.** One speaks of Newton’s method in case \( H^{(n)} \) is chosen to be

\[
H^{(n)} = \nabla^2 \mathcal{L}(\vec{\theta}^{(n)}) + \Delta H^{(n)} \tag{20}
\]

where \( \nabla^2 \mathcal{L}(\vec{\theta}) \) is the Hessian matrix of the likelihood function and the term \( \Delta H^{(n)} \) is typically selected as small as possible in order to avoid slowing convergence and to ensure that \( H^{(n)} \) is negative definite (i.e.

\[
\nabla^2 \mathcal{L}(\vec{\theta}^{(n)}) + \Delta H^{(n)} < 0.
\]

This choice of \( H^{(n)} \) is also referred to as ‘exact Hessian’.

**Quasi-Newton methods.** Any Hessian approximation which is negative definite (i.e. satisfying \( H^{(n)} < 0 \)) yields an ascent direction and guarantees convergence. Although one may choose to consider the simplest prescription \( H^{(n)} = -I \), which yields the ‘steepest ascent’ algorithm, here we have opted for the following recipe, i.e. the purely diagonal version of Newton’s method:

\[
H^{(n)}_{ii} = \nabla^2 \mathcal{L}(\vec{\theta}^{(n)}) + \Delta H^{(n)}_{ii} < 0, \quad \forall i \text{ and } H^{(n)}_{ij} = 0, \quad \forall i \neq j.
\]

**Fixed-point iteration on modified KKT conditions.** In addition to the (classes of) algorithms above, we will also consider an iterative recipe which is constructed as a fixed-point iteration on a modified version of the Karush–Kuhn–Tucker (KKT) conditions, i.e. \( F(\vec{\theta}) = 0 \) or, analogously, \( \vec{\theta} = G(\vec{\theta}) \); the iterate can, then, be made explicit by rewriting the latter as

\[
\vec{\theta}^{(n)} = G(\vec{\theta}^{(n-1)}). \tag{21}
\]

The condition above will be made explicit, for each network model, in the corresponding subsection. We also observe that this choice yields a non-standard SQP method as \( H^{(n)} \) is typically not symmetric, for our models.

**A note on convergence.** Provided that the Hessian approximation \( H^{(n)} \) is negative definite, the direction \( \Delta \vec{\theta} \) is an ascent one; as such, it is guaranteed to yield an improvement of the objective function, for a step size \( \alpha \) that is sufficiently small. The role of the backtracking line search is that of finding a step size \( \alpha \) that yields such an improvement, while making sufficient progress towards the solution. As discussed in [25], Newton’s method has local quadratic convergence, while the quasi-Newton method and the fixed-point iteration algorithm have local linear convergence.

### APPLICATIONS

Let us now apply the algorithms described in the previous section to a number of specific cases of interest. The taxonomy of the models considered in the present paper is shown in fig. 2 while the constraints defining each model are illustrated in fig. 3.

**UBCM: binary undirected graphs with given degree sequence**

Let us start by considering binary, undirected networks (BUNs). The simplest, non-trivial set of constraints is represented by the degrees of nodes: the degree of node \( i \), i.e. \( k_i(A) = \sum_{j: j \neq i} a_{ij} \), counts the number of its neighbours and coincides with the total number of 1s along the \( i \)-th row (or, equivalently, along the \( i \)-th column) of the adjacency matrix \( A \equiv \{a_{ij}\}_{i,j=1}^N \). The benchmark defined by this set of constraints is known as Undirected Binary Configuration Model (UBCM) and its Hamiltonian reads

\[
\mathcal{H}_{UBCM}(A, \vec{\theta}) = \sum_{i=1}^N \theta_i k_i(A);
\]

entropy maximization [15] [10] leads to the factorized graph probability

\[
P_{UBCM}(A | \vec{\theta}) = \prod_{i=1}^N \prod_{j=1}^N p_{ij}^{a_{ij}} (1 - p_{ij})^{1-a_{ij}} \tag{23}
\]

where \( p_{ij} = p_{ij}^{UBCM} \equiv \frac{e^{-\theta_i - \theta_j}}{1 + e^{-\theta_i - \theta_j}} \). In this case, the canonical ensemble of BUNs is the set of networks with the same number of nodes, \( N \), of the observed graph and a number of (undirected) links varying from zero to the maximum value \( \binom{N}{2} \). The argument of the problem [14] for the specific network \( A^* \) becomes

\[
\mathcal{L}_{UBCM}(\vec{\theta}) = -\sum_{i=1}^N \theta_i k_i(A^*) - \sum_{i=1}^N \sum_{j=1}^N \ln \left[ 1 + e^{-\theta_i - \theta_j} \right] \tag{24}
\]

whose first-order optimality conditions read

\[
\frac{\partial \mathcal{L}_{UBCM}(\vec{\theta})}{\partial \theta_i} = 0, \quad \forall i
\]

\[
\frac{\partial \mathcal{L}_{UBCM}(\vec{\theta})}{\partial \theta_j} = 0, \quad \forall j
\]

\[
\frac{\partial \mathcal{L}_{UBCM}(\vec{\theta})}{\partial \theta_{ij}} = 0, \quad \forall i \neq j
\]
\[ \nabla_{\theta_i} \mathcal{L}_{\text{UBCM}} = -k_i(A^*) + \sum_{j=1 \atop j \neq i}^{N} \frac{e^{-\theta_j - \theta_i}}{1 + e^{-\theta_i - \theta_j}} \]
\[ = -k_i(A^*) + \sum_{j=1 \atop j \neq i}^{N} P_{ij}^{\text{UBCM}} \]
\[ = -k_i(A^*) + \langle k_i \rangle = 0, \quad i = 1 \ldots N. \tag{25} \]

Resolution of the UBCM. Newton’s and the quasi-Newton method can be easily implemented via the recipe defined in eq. (18) (see Appendix A for the definition of the UBCM Hessian).

The explicit definition of the fixed-point recipe, instead, requires a preliminary observation, i.e. that the system of equations embodying the UBCM first-order optimality conditions can be re-written as follows
\[ e^{-\theta_i} = \frac{k_i(A^*)}{\sum_{j=1 \atop j \neq i}^{N} \left( \frac{e^{-\theta_j}}{1 + e^{-\theta_i - \theta_j}} \right)}, \quad i = 1 \ldots N \tag{26} \]
i.e. as a set of consistency equations. The observation that the term \( e^{-\theta_i} \) appears on both sides of the equation corresponding to the \( i \)-th constraint suggests an iterative recipe to solve such a system, i.e.
\[ \theta_i^{(n)} = -\ln \left[ \frac{k_i(A^*)}{\sum_{j=1 \atop j \neq i}^{N} \left( \frac{e^{-\theta_j}}{1 + e^{-\theta_i - \theta_j}} \right)^{n}} \right], \quad i = 1 \ldots N \tag{27} \]
originally proposed in [22] and further refined in [23].

The identification \( p_{ij}^{\text{UBCM}} = \frac{e^{-\theta_{ij}^{(n)}}}{1 + e^{-\theta_{ij}^{(n-1)} - \theta_{ij}^{(n-1)}}}, \quad \forall i < j \) allows the probability coefficients defining the UBCM to be numerically determined.

As any other iterative recipe, the one proposed above needs to be initialized as well. To this aim, we have tested three different sets of initial values: the first one is defined by the position \( \theta_i^{(0)} = -\ln \left[ \frac{k_i(A^*)}{\sqrt{2N}} \right], \quad \forall i \) - usually, a good approximation of the solution of the system of equations in (25), in the ‘sparse case’ (i.e. whenever \( p_{ij}^{\text{UBCM}} \approx e^{-\theta_{ij}}, i < j \) [27]; the second one is a variant of the position above, reading \( \theta_i^{(0)} = -\ln \left[ \frac{k(A^*)}{\sqrt{N}} \right], \quad \forall i \); the third one, instead, prescribes to randomly draw the value of each parameter from a uniform distribution with support on the unit interval, i.e. \( \theta_i^{(0)} \sim U(0, 1), \forall i. \)

Reducing the dimensionality of the problem. The problem defining the UBCM can be further simplified by noticing that nodes with the same degree, say \( k \), can be assigned the same value of the multiplier \( \theta \) [24] - a result resting upon the observation that any value \( k_i(A^*) \) must match the sum of monotonic, increasing functions. This translates into the possibility of rewriting \( \mathcal{L}_{\text{UBCM}}(\theta) \) in a ‘reduced’ fashion, as
\[ \mathcal{L}_{\text{UBCM}}^{\text{reduced}}(\vec{\theta}) = - \sum_{k} f(k) \theta_k k(A^*) \]
\[ - \sum_{k} \sum_{k' \leq k} f(k) [f(k') - \delta_{kk'}] \cdot \ln \left[ 1 + e^{-\theta_k - \theta_{k'}} \right] \tag{28} \]
where the sums run over the distinct values of the degrees and \( f(k) \) counts the number of nodes whose degree is \( k \). Rewriting the problem with respect to the set \( \{ \theta_k \}_k \) leads one to recover simplified versions of the three algorithms considered here: Newton’s and the quasi-Newton methods can, now, be solved via a ‘reduced’ version of eqs. (18) (since both the dimension of the gradient and the order of the Hessian matrix of the likelihood function are, now, less than \( N \)), while the iterative recipe defined in (26) can be rewritten in terms of the ‘non-degenerate’ degrees, as
\[ \theta_k^{(n)} = -\ln \left[ \frac{k(A^*)}{\sum_{k'|k} [f(k') - \delta_{kk'}] \left( \frac{e^{-\theta_{k'}}}{1 + e^{-\theta_k - \theta_{k'}}} \right)} \right] \tag{29} \]
\[ \forall k, \text{ where, at the denominator, the self-contribution (i.e. the probability that a node links to itself) has been explicitly excluded.} \]

Performance testing. The performance of the three algorithms, considered in the present paper, to solve the reduced version of eqs. (25), has been tested on a bunch of real-world networks. The latter ones span a wide variety of systems, including natural, financial and technological ones. In particular, we have considered the synthetic network of the worm \( C. \ Elegans \) [28], the network of the largest US airports [29], the protein-protein interaction network of the bacterium \( H. \ Pylori \) [30], Internet at the level of Autonomous Systems [31] and eight daily snapshots of the so-called Bitcoin Lightning Network [32], chosen throughout its entire history. Before commenting on the results of our numerical exercises, let us, first, describe how the latter ones have been carried out.

The accuracy of each algorithm in reproducing the constraints defining the UBCM has been quantified via the
maximum absolute error metrics, defined, in a perfectly general fashion, as \( \max_i \{|C_i^* - (C_i)|\}_i \) where \( C_i^* \) is the empirical value of the \( i \)-th constraint, \( C_i \). Naturally, in the UBCM case, \( C_i = k_i \), \( \forall i \) and the aforementioned error score becomes

\[
\text{MADE} = \max \{|k_i^* - \langle k_i \rangle|\}_i \quad (30)
\]

(the acronym standing for Maximum Absolute Degree Error). Equivalently, it is the infinite norm of the difference between the vector of the empirical values of the constraints and that of their expected values.

For each algorithm, we have considered three different stopping criteria: the first one puts a condition on the Euclidean norm of the gradient of the likelihood function, i.e.

\[
\|\nabla \mathcal{L}(\hat{\theta})\|_2 = \sqrt{\sum_{i=1}^N (\nabla_i \mathcal{L}(\hat{\theta}))^2} \leq 10^{-8}; \quad (31)
\]

the second one puts a condition on the Euclidean norm of the vector of differences between the values of the parameters at subsequent iterations, i.e.

\[
\|\Delta \tilde{\theta}\|_2 = \sqrt{\sum_{i=1}^N (\Delta \tilde{\theta}_i)^2} \leq 10^{-8}; \quad (32)
\]

the third one concerns the maximum number of iterations: after 1000 steps, any of the three algorithms stops.

The results about the performance of our three algorithms are reported in Table I. Overall, all recipes perform very satisfactorily, being accurate, fast and scalable; moreover, all algorithms stop either because the condition on the norm of the likelihood is satisfied or because the condition on the norm of the vector of parameters is satisfied.

For what concerns accuracy, the largest maximum error per method spans an interval (across all configurations) that amounts at \( 10^{-10} \leq \text{MADE}_{\text{reduced Newton}} \leq 10^{-6} \). The best results are obtained by the quasi-Newton method and the fixed-point one; however, the former is faster and the latter is more accurate.

It is evident that the two most accurate methods are Newton’s and the fixed-point ones; for what concerns speed, the fastest method is the fixed-point one (although Newton’s one approximately requires the same amount of time on each specific configuration). Only the results corresponding to the best choice of initial conditions are reported.

| Algorithm | N | L (ms) | c | \( c_r \) | MADE | Time (s) | MADE | Time (s) | MADE | Time (s) |
|-----------|---|-------|---|---------|-------|-------|-------|-------|-------|-------|
| C. Elegans (nn) | 265 | 1879 | \( \approx 5 \cdot 10^{-2} \) | \( \approx 1.5 \cdot 10^{-1} \) | \( \approx 8.1 \cdot 10^{-5} \) | \( \approx 1 \cdot 10^{-6} \) | \( \approx 0.005 \) | \( \approx 0.03 \) | \( \approx 5 \cdot 10^{-8} \) | \( \approx 0.004 \) |
| US airports | 500 | 2980 | \( \approx 2 \cdot 10^{-2} \) | \( \approx 1.3 \cdot 10^{-1} \) | \( \approx 8.9 \cdot 10^{-5} \) | \( \approx 1.2 \cdot 10^{-6} \) | \( \approx 0.008 \) | \( \approx 0.04 \) | \( \approx 2.5 \cdot 10^{-8} \) | \( \approx 0.005 \) |
| H. Pylori (pp) | 732 | 1465 | \( \approx 5 \cdot 10^{-3} \) | \( \approx 4.5 \cdot 10^{-4} \) | \( \approx 1.3 \cdot 10^{-8} \) | \( \approx 6.7 \cdot 10^{-11} \) | \( \approx 0.005 \) | \( \approx 0.07 \) | \( \approx 7 \cdot 10^{-8} \) | \( \approx 0.014 \) |
| Internet (AS) | 11174 | 23409 | \( \approx 4 \cdot 10^{-4} \) | \( \approx 1 \cdot 10^{-2} \) | \( \approx 4.1 \cdot 10^{-11} \) | \( \approx 5 \cdot 10^{-14} \) | \( \approx 0.03 \) | \( \approx 0.10 \) | \( \approx 2 \cdot 10^{-4} \) | \( \approx 0.005 \) |

| Algorithm | BLN 24-01-18 | BLN 25-02-18 | BLN 30-03-18 | BLN 13-07-18 | BLN 19-12-18 | BLN 30-01-19 | BLN 01-03-19 | BLN 17-07-19 |
|-----------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Newton | 94 | 152 | \( \approx 3 \cdot 10^{-4} \) | \( \approx 1.5 \cdot 10^{-3} \) | \( \approx 1 \cdot 10^{-5} \) | \( \approx 3 \cdot 10^{-6} \) | \( \approx 0.01 \) | \( \approx 1 \cdot 10^{-4} \) | \( \approx 0.03 \) | \( \approx 5 \cdot 10^{-8} \) | \( \approx 0.002 \) |
| Quasi-Newton | 499 | 1011 | \( \approx 8 \cdot 10^{-3} \) | \( \approx 6.4 \cdot 10^{-2} \) | \( \approx 1.6 \cdot 10^{-5} \) | \( \approx 9.1 \cdot 10^{-11} \) | \( \approx 0.02 \) | \( \approx 7.3 \cdot 10^{-8} \) | \( \approx 0.004 \) |
| Fixed-point | 1012 | 2952 | \( \approx 5 \cdot 10^{-4} \) | \( \approx 5.2 \cdot 10^{-2} \) | \( \approx 3.4 \cdot 10^{-19} \) | \( \approx 2 \cdot 10^{-30} \) | \( \approx 0.03 \) | \( \approx 1.4 \cdot 10^{-19} \) | \( \approx 0.005 \) |

TABLE I. Performance of Newton’s, quasi-Newton and the fixed-point algorithm to solve the reduced system of equations defining the UBCM, on a set of real-world BUNs (of which basic statistics as the total number of nodes, \( N \), the total number of links, \( L \), and the connectance, \( c = 2L/N(N-1) \), are provided). All algorithms stop either because the condition \( \|\nabla \mathcal{L}(\hat{\theta})\|_2 \leq 10^{-8} \) is satisfied or because the condition \( \|\Delta \tilde{\theta}\|_2 \leq 10^{-8} \) is satisfied. For what concerns accuracy, the two most accurate methods are Newton’s and the fixed-point ones; for what concerns speed, the fastest method is the fixed-point one (although Newton’s one approximately requires the same amount of time on each specific configuration). Only the results corresponding to the best choice of initial conditions are reported.
that actually need to be solved.

While reducibility can be easily quantified a posteriori, e.g. by calculating the coefficient of reduction, $c_r$, defined as the ratio between the number of equations that survive to reduction and the number of equations defining the original problem (hence, the smaller the better), providing an exhaustive list of the aforementioned factors a priori is much more difficult.

In the case of the UBCM, $c_v$ is defined as the number of different degrees divided by the total number of nodes; one may, thus, argue that reducibility is affected by the heterogeneity of the degree distribution; upon considering that the latter can be quantified by computing the coefficient of variation (defined as $c_v = s/m$, where $s$ and $m$ are, respectively, the standard deviation and the mean of the degree distribution of the network at hand), one may derive a simple rule of thumb: a larger coefficient of variation (pointing out a larger heterogeneity of the degree distribution) leads to a larger coefficient of reduction and a larger amount of time for convergence will be required. Notice that even if the degree distribution is narrow, outliers (e.g. hubs) may still play a role, forcing the corresponding parameters to assume either very large or very small values - hence, slowing down the entire convergence process.

In this sense, scalability is the result of a (non-trivial) interplay between size and reducibility. Let us take a look at Table II Internet is the most reducible network of our basket, although being the largest in size, while the neural network of C. Elegans is one of the least reducible networks of our basket, although being the second smallest one; as a consequence, the actual number of equations defining the UBCM on C. Elegans is $\approx 30$ while the actual number of equations defining the UBCM on Internet is $\approx 100$ - whence the larger amount of time to solve the latter. Remarkably, the time required by our recipes to ensure that the largest system of equations converges to the solution ranges from thousandths to tenths of seconds.

As a last comment, we would like to stress that, unlike several popular approximations as the Chung-Lu one, the generic coefficient $p_{ij}^{UBCM}$ always represents a proper probability, in turn implying that eq. (23) also provides us with a recipe to sample the canonical ensemble of BUNs, under the UBCM. Notice that the factorization of the graph probability $P_{UBCM}(\mathbf{A}|\vec{\theta})$ greatly simplifies the entire procedure, allowing a single graph to be sampled by implementing the Bernoulli trial

$$a_{ij} = \begin{cases} 0 & 1 - p_{ij}^{UBCM} \\ 1 & p_{ij}^{UBCM} \end{cases}$$

for each (undirected) pair of nodes, in either a sequential or a parallel fashion. The sampling process, whose computational complexity amounts at $O(N^2)$, can be repeated to generate as many configurations as desired. The pseudo-code for explicitly sampling the UBCM ensemble is summed up by Algorithm 1.

We explicitly acknowledge the existence of the algorithm proposed in [33] for sampling binary, undirected networks from the Chung-Lu model (i.e. the ‘sparse case’ approximation of the UBCM), a recipe that is applicable whenever the condition $p_{ij}^{CL} = k_i k_j / N^2 < 1$, $\forall i < j$ is verified. As explicitly acknowledged by the authors of [32], however, it does not hold in several cases of interest: an example of paramount importance is provided by sparse networks whose degree distribution is scale-free. In such cases, $k_{max} \sim N^{\gamma}$; hence, the hubs establish a connection with probability $p_{ij}^{CL} \sim N \frac{\gamma}{N - N^{\gamma-2}}$ that becomes larger than 1 when $2 < \gamma \leq 3$ and diverges for $\gamma \to 2$, thus leading to a strong violation of the requirement above.

**Algorithm 1** Sampling the UBCM ensemble

1: for $m = 1 \ldots |E|$ do
2: \hspace{1cm} $\mathbf{A} = \mathbf{0}$;
3: for $i = 1 \ldots N$ do
4: \hspace{1cm} for $j = 1 \ldots N$ and $j < i$ do
5: \hspace{2cm} if RandomUniform$[0,1] \leq p_{ij}^{UBCM}$ then
6: \hspace{3cm} $a_{ij} = a_{ji} = 1$;
7: \hspace{2cm} else
8: \hspace{3cm} $a_{ij} = a_{ji} = 0$;
9: \hspace{2cm} end if
10: \hspace{1cm} end for
11: \hspace{1cm} end for
12: Ensemble$[m] = \mathbf{A}$;
13: end for

**DBCM: binary directed graphs**

with given in-degree and out-degree sequences

Let us now move to consider binary, directed networks (BDNs). In this case, the simplest, non-trivial set of constraints is represented by the in-degrees and the out-degrees of nodes, where $k_i^{in}(\mathbf{A}) = \sum_{j(\neq i)=1}^N a_{ij}$ counts the number of nodes ‘pointing’ to node $i$ and $k_i^{out}(\mathbf{A}) = \sum_{i(\neq j)=1}^N a_{ij}$ counts the number of nodes $i$ ‘points’ to. The benchmark defined by this set of constraints is known as Directed Binary Configuration Model (DBCM) whose Hamiltonian reads

$$H_{DBCM}(\mathbf{A}, \vec{\alpha}, \vec{\beta}) = \sum_{i=1}^N [\alpha_i k_i^{out}(\mathbf{A}) + \beta_i k_i^{in}(\mathbf{A})]$$

as in the undirected case, entropy maximization \cite{15, 16} leads to a factorized probability distribution, i.e.
where \( p_{ij} = p_{ij}^{\text{DBCM}} = \frac{e^{-\alpha_i - \beta_j}}{1 + e^{-\alpha_i - \beta_j}} \). The canonical ensemble of BDNs is, now, the set of networks with the same number of nodes, \( N \), of the observed graph and a number of (directed) links varying from zero to the maximum value \( N(N-1) \). The argument of the problem \([14]\) for the specific network \( A^* \) becomes

\[
P_{\text{DBCM}}(A^|\alpha, \beta) = \prod_{i=1}^{N} \prod_{j=1}^{N} p_{ij}^{\alpha \beta} (1 - p_{ij})^{1-\alpha \beta} \tag{35}
\]

The fixed-point recipe for solving the system of equations embodying the DBCM first-order optimality conditions can, instead, be re-written in the usual iterative fashion as follows

\[
\alpha_i^{(n)} = -\ln \left[ \frac{k^{\text{out}}_i(A^*)}{\sum_{j=1}^{N} \frac{e^{-\alpha_i - \beta_j}}{1 + e^{-\alpha_i - \beta_j}}} \right], \quad i = 1 \ldots N
\]

\[
\beta_i^{(n)} = -\ln \left[ \frac{k^{\text{in}}_i(A^*)}{\sum_{j=1}^{N} \frac{e^{-\alpha_i - \beta_j}}{1 + e^{-\alpha_i - \beta_j}}} \right], \quad i = 1 \ldots N
\tag{39}
\]

Analogously to the undirected case, the initialization of this recipe has been implemented in three different ways. The first one reads \( \alpha_i^{(0)} = -\ln \left[ \frac{k^{\text{out}}_i(A^*)}{\sqrt{L}} \right], \quad i = 1 \ldots N \) and \( \beta_i^{(0)} = -\ln \left[ \frac{k^{\text{in}}_i(A^*)}{\sqrt{N}} \right], \quad i = 1 \ldots N \) and represents a good approximation to the solution of the system of equations defining the DBCM in the ‘sparse case’ (i.e. whenever \( p_{ij}^{\text{DBCM}} \simeq e^{-\alpha_i - \beta_j} \)); the second one is a variant of the position above, reading \( \alpha_i^{(0)} = -\ln \left[ \frac{k^{\text{out}}_i(A^*)}{\sqrt{N}} \right], \quad i = 1 \ldots N \) and \( \beta_i^{(0)} = -\ln \left[ \frac{k^{\text{in}}_i(A^*)}{\sqrt{N}} \right], i = 1 \ldots N \); the third one, instead, prescribes to randomly draw the value of each parameter from a uniform distribution defined on the unit interval, i.e. \( \alpha_i^{(0)} \sim U(0,1), \forall i \) and \( \beta_i^{(0)} \sim U(0,1), \forall i \). As for the UBCM, the identification \( p_{ij}^{\text{DBCM}} \equiv \frac{e^{-\alpha_i - \beta_j}}{1 + e^{-\alpha_i - \beta_j}}, \forall i \neq j \) allows the probability coefficients defining the DBCM to be numerically determined.

**Reducing the dimensionality of the problem.** As for the UBCM, we can define a ‘reduced’ version of the DBCM likelihood, accounting only for the distinct (pairs of) values of the degrees. By defining \( k^{\text{out}} \equiv k \) and \( k^{\text{in}} \equiv h \), in order to simplify the formalism, the reduced DBCM recipe reads

\[
\mathcal{L}_{\text{DBCM-reduced}}(\bar{\theta}) = -\sum_{k} \sum_{h} n(k,h) [\alpha_k,h_k(A^*) + \beta_k,k h(A^*)]
\]

\[
-\sum_{k,h} \sum_{k',h'} n(k,h) [n(k',h') - \delta_{kk'} \delta_{hh'}] \cdot \ln \left[ 1 + e^{-\alpha_k,h - \beta_{k',h'}} \right]
\tag{40}
\]

The implementation of the algorithms considered here must be modified in a way that is analogous to the one already described for the UBCM. In particular, the fixed-point recipe for the DBCM can be re-written by assigning to the nodes with the same out- and in-degrees \((k,h)\) the same pair of values \((\alpha, \beta)\), i.e. as

Resolution of the DBCM. Newton’s and the quasi-Newton method can be easily implemented via the recipe defined in eq. \([18]\) (see Appendix A for the definition of the DBCM Hessian).
\[ a^{(n)}_{k,h} = -\ln \left( \frac{k(A^*)}{\sum k',h' \left[ n(k',h') - \delta_{kk'} \delta_{hh'} \right]} \right), \quad \forall k, h \] 
\[ \beta^{(n)}_{k,h} = -\ln \left( \frac{h(A^*)}{\sum k',h' \left[ n(k',h') - \delta_{kk'} \delta_{hh'} \right]} \right), \quad \forall k, h \] 

where the sums, now, run over the distinct values of the out- and in-degrees, \( n(k, h) \) is the number of nodes whose out-degree is \( k \) and whose in-degree is \( h \), and, as usual, the last term at the denominator excludes the self-contribution (i.e. the probability that a node links to itself).

**Performance testing.** As for the UBCM, the performance of the three algorithms in solving the reduced version of eqs. (37) and (38) has been tested on a bunch of real-world networks. The latter ones span economic, financial and social networks. In particular, we have considered the World Trade Web (WTW) during the decade 1992-2002 [33], a pair of snapshots of the Bitcoin User Network at the weekly time scale (the first day of those weeks being 13-02-12 and 27-04-15, respectively) [35] and of the corresponding largest weakly connected component (whose size is, respectively, \( \approx 65\% \) and \( \approx 90\% \) of the full network size) and a snapshot of the semantic network concerning the Twitter discussion about the Covid-19 pandemics (more precisely: it is the network of re-tweets of the (online) Italian debate about Covid-19, collected in the period 21st February - 20th April 2020) [36]. Before commenting on the results of our numerical exercises, let us, first, describe how the latter ones have been carried out.

The accuracy of each algorithm in reproducing the constraints defining the DBCM has been quantified via the maximum absolute error metrics that, in this case, reads

\[ \text{MADE} = \max_i \left\{ |k_i^* - \langle k_i \rangle|, |h_i^* - \langle h_i \rangle| \right\} \] 

and accounts for the presence of two different degrees per node. As for the UBCM, it is the infinite norm of the difference between the vector of the empirical values of the constraints and that of their expected values.

The three different ‘stop criteria’ we have considered match the ones adopted for analysing the undirected case and consist in a condition on the Euclidean norm of the gradient of the likelihood function, i.e. \( \| \nabla \mathcal{L}(\theta) \|_2 \leq 10^{-8} \), in a condition on the Euclidean norm of the vector of differences between the values of the parameters at subsequent iterations, i.e. \( \| \Delta \bar{\theta} \|_2 \leq 10^{-8} \), and in a condition on the maximum number of iterations: after 10000 steps, any of the three algorithms stops.

The results about the performance of our three algorithms are reported in Table I. Overall, all recipes perform very satisfactorily, being accurate, fast and scalable; however, while Newton’s and the quasi-Newton methods stop because the condition on the norm of the likelihood is satisfied, the fixed-point recipe is always found to satisfy the limit condition on the number of steps (i.e. it runs for 10000 steps and, then, stops).

Let us start by commenting the results concerning the WTW. For what concerns accuracy, the largest maximum error per method spans an interval, across all configurations, that amounts at \( \text{MADE}_{\text{reduced}} \approx 10^{-9} \), \( 10^{-10} \approx \text{MADE}_{\text{quasi-Newton}} \approx 10^{-8} \) and \( \text{MADE}_{\text{fixed-point}} \approx 10^{-2} \). By looking at each specific network, it is evident that the most accurate method is systematically the quasi-Newton one.

For what concerns speed, the amount of time required by each method to achieve convergence spans an interval (across all configurations) that is \( 0.5 \leq T_{\text{Newton}} \leq 13 \), \( 0.21 \leq T_{\text{quasi-Newton}} \leq 0.33 \) and \( 2.5 \leq T_{\text{fixed-point}} \leq 4.4 \) (time is measured in seconds). The fastest method is the quasi-Newton one, followed by Newton’s method and the fixed-point recipe. The latter is the slowest method since it always reaches the limit of 10000 steps while the other two competing ones stop after few iterations. Appreciable differences in the speed of convergence of any method, caused by the choice of a particular set of initial conditions, are not observed.

The observations above hold true when the WTW is considered. The picture changes when very large networks, as Bitcoin and Twitter, are considered. First, let us notice that Bitcoin and Twitter ‘behave’ as the undirected version of Internet considered to solve the UBCM, i.e. they are very redundant, hosting many nodes with the same out- and in-degrees (in fact, the coefficient of reduction, \( c_r \), is, now, defined as the number of differ-
MADE in 1989.

\[ \sum_{\text{quadrant}} \equiv \begin{cases} 0 & \text{if } p \leq P_{\text{PCMC}} \\ 1 & \text{if } p > P_{\text{PCMC}} \end{cases} \]

As a last comment, we would like to stress that, as in the undirected case, the test coefficient \( p \) is defined as the number of times that the undirected graph of the graph probability \( P_{\text{PCMC}}(G) \) is sampled by the Bernoulli trial, allowing a single graph to be sampled by implementing the Bernoulli trial for each directed pair of nodes in either a sequential or a parallel fashion. The sampling process, whose calculation is (very) time demanding due to the need to sample the directed configuration, is performed by generating random numbers as desired. The pseudocode for explicitly sampling the DCM is

\[ a_k = \begin{cases} 0 & \text{if } P_{\text{PCMC}} \\ 1 & \text{if } P_{\text{PCMC}} \end{cases} \]

For our binary, directed networks, in fact, the resolution in a reasonable amount of time; on the other hand, the reducibility of the directed configurations considered here is what guarantees their solution. For instance, a good proxy of the reliability of the directed configurations considered here is provided by their connectivity. For instance, the reliability is the fixed-point one. Only the results corresponding to the best choice of initial conditions are reported. Newton's method and its quasi-Newton counterpart are widely used in practice. However, while Newton's method is generally considered the most accurate method, for binary directed networks, in fact, the quasi-Newton method becomes the fixed-point method. In order to understand the reason underlying this result, we need to consider that both Newton's and quasi-Newton methods are iterative; the former, however, requires a lot of memory for the step-wise update, while the latter requires a lot of memory for the update, but is more memory-efficient when the network size is large. On the one hand, this feature, common to the very high-dimensional systems, guarantees that the quasi-Newton method is more suitable for large networks. On the other hand, this feature, common to the very high-dimensional systems, guarantees that the quasi-Newton method is more suitable for large networks.

As a last comment, we would like to stress that, as in the undirected case, the test coefficient \( p \) is defined as the number of times that the undirected graph of the graph probability \( P_{\text{PCMC}}(G) \) is sampled by the Bernoulli trial, allowing a single graph to be sampled by implementing the Bernoulli trial for each directed pair of nodes in either a sequential or a parallel fashion. The sampling process, whose calculation is (very) time demanding due to the need to sample the directed configuration, is performed by generating random numbers as desired. The pseudocode for explicitly sampling the DCM is

\[ a_k = \begin{cases} 0 & \text{if } P_{\text{PCMC}} \\ 1 & \text{if } P_{\text{PCMC}} \end{cases} \]
Algorithm 2 Sampling the DBCM ensemble

1: for $m = 1 \ldots |E|$ do
2: $A = 0$;
3: for $i = 1 \ldots N$ do
4: for $j = 1 \ldots N$ and $j \neq i$ do
5: if RandomUniform[0, 1] $\leq p_{ij}^{\text{DBCM}}$ then
6: $a_{ij} = 1$;
7: else
8: $a_{ij} = 0$;
9: end if
10: end for
11: end for
12: Ensemble[$m$] = $A$;
13: end for

BiCM: bipartite binary undirected graphs with given degree sequences

So far, we have considered monopartite networks. However, the algorithm we have described for solving the DBCM can be adapted, with little effort, to solve a null model designed for bipartite, binary, undirected networks (BiBUNs), i.e. the so-called Bipartite Configuration Model (BiCM) [27]. These networks are defined by two distinct layers (say, $\top$ and $\bot$) and obey the rule that links can exist only between (and not within) layers: for this reason, they can be compactly described via a biadjacency matrix $B \equiv \{b_{\alpha\beta}\}$, whose generic entry $b_{\alpha\beta}$ is 1 if node $i$ belonging to layer $\bot$ is linked to node $\alpha$ belonging to layer $\top$ and 0 otherwise. The constraints defining the BiCM are represented by the degree sequences $\{k_i\}_{i \in \bot}$ and $\{d_\alpha\}_{\alpha \in \top}$ where $k_i = \sum_{\alpha \in \top} b_{i\alpha}$ counts the neighbors of node $i$ (belonging to layer $\top$) and $d_\alpha = \sum_{i \in \bot} b_{i\alpha}$ counts the neighbors of node $\alpha$ (belonging to layer $\bot$).

Analogously to the DBCM case,

$$P(B|\gamma, \beta) = \prod_{i \in \bot} \prod_{\alpha \in \top} p_{\alpha\beta}^{b_{i\alpha}} (1 - p_{\alpha\beta})^{1 - b_{i\alpha}}$$  \hspace{1cm} (45)

where $p_{\alpha\beta} = p_{\alpha\beta}^{\text{BiCM}} \equiv \frac{e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}{1 + e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}$. The canonical ensemble of BiBUNs includes all networks with, say, $N$ nodes on one layer, $M$ nodes on the other layer and a number of links (connecting nodes of different layers) ranging from zero to the maximum value $N \cdot M$.

The BiCM likelihood function reads

$$\mathcal{L}_{\text{BiCM}}(\gamma, \beta) = -\sum_{i \in \bot} \gamma_i k_i(B^*) - \sum_{\alpha \in \top} \beta_\alpha d_\alpha(B^*)$$
$$- \sum_{i \in \bot} \sum_{\alpha \in \top} \ln [1 + e^{-\gamma_{i\alpha} - \beta_{i\alpha}}]$$ \hspace{1cm} (46)

whose first-order optimality conditions read

$$\nabla_{\gamma_i} \mathcal{L}_{\text{BiCM}} = -k_i(B^*) + \sum_{\alpha \in \top} \frac{e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}{1 + e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}, \quad i \in \bot$$
$$\nabla_{\beta_\alpha} \mathcal{L}_{\text{BiCM}} = -d_\alpha(B^*) + \sum_{i \in \bot} \frac{e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}{1 + e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}, \quad \alpha \in \top.$$ \hspace{1cm} (47)

**Resolution of the BiCM.** As for the DBCM case, Newton’s and the quasi-Newton methods can be implemented by adapting the recipe defined in eq. [18] to the bipartite case (see Appendix A for the definition of the BiCM Hessian).

As for the DBCM, the fixed-point recipe for the BiCM can be re-written in the usual iterative fashion as follows

$$\gamma_i^{(n)} = -\ln \left[ \frac{k_i(B^*)}{\sum_{\alpha \in \top} \left( \frac{e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}{1 + e^{-\gamma_{i\alpha} - \beta_{i\alpha}}} \right)} \right], \quad \forall i \in \bot$$
$$\beta_\alpha^{(n)} = -\ln \left[ \frac{d_\alpha(B^*)}{\sum_{i \in \bot} \left( \frac{e^{-\gamma_{i\alpha} - \beta_{i\alpha}}}{1 + e^{-\gamma_{i\alpha} - \beta_{i\alpha}}} \right)} \right], \quad \forall \alpha \in \top.$$ \hspace{1cm} (48)

and the initialization is similar as well: in fact, we can employ the value of the solution of the BiCM in the sparse case, i.e. $\gamma_i^{(0)} = -\ln \left[ \frac{k_i(B^*)}{\sqrt{L}} \right], \forall i \in \bot$ and $\beta_\alpha^{(0)} = -\ln \left[ \frac{d_\alpha(B^*)}{\sqrt{L}} \right], \forall \alpha \in \top$ (only this set of initial conditions has been employed to analyse the bipartite case).

**Reducing the dimensionality of the problem.** Exactly as for the DBCM case, the presence of nodes with the same degree, on the same layer, leads to the appearance of identical equations in the system above; hence, the computation of the solutions can be sped up by writing

$$\gamma_k^{(n)} = -\ln \left[ \frac{k_k(B^*)}{\sum_d g(d) \left( \frac{e^{-\gamma_k d_{k,d}}}{1 + e^{-\gamma_k d_{k,d}}} \right)} \right], \quad \forall k$$
$$\beta_d^{(n)} = -\ln \left[ \frac{d_B(B^*)}{\sum_k f(k) \left( \frac{e^{-\gamma_k d_{k,d}}}{1 + e^{-\gamma_k d_{k,d}}} \right)} \right], \quad \forall d.$$ \hspace{1cm} (49)

where $f(k)$ is the number of nodes, belonging to layer $\bot$, whose degree is $k$ and $g(d)$ is the number of nodes, belonging to layer $\top$, whose degree is $d$. 


### Performance testing.

The performance of the three algorithms in solving eqs. \((19)\) has been tested on 16 snapshot of the bipartite, binary, undirected version of the WTW, gathering the country-product export relationships across the years 1995-2010 \([37]\). Before commenting on the results of our numerical exercises, let us, first, describe how the latter ones have been carried out.

The accuracy of each algorithm in reproducing the constraints defining the BiCM has been quantified via the maximum absolute error metrics that, now, reads

\[
\text{MADE} = \max\{|k_1^1 - \langle k_1 \rangle|, \ldots, |k_N - \langle k_N \rangle|, |d_1^1 - \langle d_1 \rangle|, \ldots, |d_M - \langle d_M \rangle|\}
\]

(50)
to account for the degrees of nodes on both layers.

The three different ‘stop criteria’ match the ones adopted for analysing the UBCM and the DBCM and consist in a condition on the Euclidean norm of the gradient of the likelihood function, i.e. \(||\nabla \mathcal{L}(\hat{\theta})||_2 \leq 10^{-10}\) in a condition on the Euclidean norm of the vector of differences between the values of the parameters at subsequent iterations, i.e. \(||\Delta \hat{\theta}||_2 \leq 10^{-10}\), and a condition on the maximum number of iterations: after 1000 steps, any of the three algorithms stops.

The results about the performance of our three algorithms are reported in Table III. Overall, all recipes are accurate, fast and scalable; all methods stop because the condition on the norm of the likelihood is satisfied.

|       | Newton | Quasi-Newton | Fixed-point |
|-------|--------|--------------|-------------|
| WTW 95 | 1277  | 18947 | 1.10^{-13} | 1.30^{-13} | 0.0022 | 3.00.10^{-6} | 0.012 | 7.70.10^{-6} | 0.012 |
| WTW 96 | 1277  | 19934 | 1.10^{-13} | 1.30^{-13} | 0.0023 | 1.50.10^{-5} | 0.014 | 0.24.10^{-3} | 0.014 |
| WTW 97 | 1277  | 20222 | 1.10^{-13} | 1.30^{-13} | 0.0025 | 3.50.10^{-8} | 0.02 | 2.40.10^{-4} | 0.015 |
| WTW 98 | 20614 | 1.10^{-13} | 1.40^{-13} | 2.80.10^{-13} | 0.0024 | 1.20.10^{-6} | 0.015 | 1.80.10^{-6} | 0.015 |
| WTW 99 | 20949 | 1.40^{-13} | 1.40^{-13} | 1.30^{-13} | 0.0023 | 2.80.10^{-3} | 0.012 | 2.10.10^{-4} | 0.019 |
| WTW 100| 21257 | 1.10^{-13} | 1.40^{-13} | 1.30^{-13} | 0.0025 | 1.30.10^{-6} | 0.016 | 2.80.10^{-6} | 0.018 |
| WTW 101| 21326 | 1.10^{-13} | 1.40^{-13} | 1.70^{-13} | 0.0023 | 3.40.10^{-6} | 0.015 | 2.50.10^{-6} | 0.015 |
| WTW 102| 21333 | 1.10^{-13} | 1.40^{-13} | 1.70^{-13} | 0.0024 | 4.10^{-10} | 0.018 | 2.10.10^{-4} | 0.016 |
| WTW 103| 21330 | 1.10^{-13} | 1.30^{-13} | 2.80.10^{-13} | 0.0023 | 1.10.10^{-6} | 0.015 | 4.30.10^{-6} | 0.014 |
| WTW 104| 21479 | 1.30^{-13} | 1.30^{-13} | 1.70^{-13} | 0.0024 | 2.20.10^{-6} | 0.018 | 2.10.10^{-4} | 0.019 |
| WTW 105| 21814 | 1.10^{-13} | 1.40^{-13} | 2.30^{-13} | 0.0024 | 2.20.10^{-6} | 0.013 | 2.30.10^{-4} | 0.027 |
| WTW 106| 21945 | 1.30^{-13} | 1.30^{-13} | 2.30^{-13} | 0.0024 | 1.30.10^{-6} | 0.016 | 2.20.10^{-4} | 0.012 |
| WTW 107| 22036 | 1.30^{-13} | 1.40^{-13} | 2.30^{-13} | 0.0024 | 2.00.10^{-10} | 0.017 | 2.10.10^{-4} | 0.023 |
| WTW 108| 21889 | 1.30^{-13} | 1.30^{-13} | 1.10^{-13} | 0.0024 | 1.50.10^{-6} | 0.017 | 2.50.10^{-4} | 0.024 |
| WTW 109| 21621 | 1.30^{-13} | 1.30^{-13} | 2.30^{-13} | 0.0025 | 2.10.10^{-6} | 0.021 | 2.40.10^{-4} | 0.018 |
| WTW 10 | 21010 | 1.30^{-13} | 1.30^{-13} | 2.30^{-13} | 0.0022 | 1.60.10^{-10} | 0.015 | 2.60.10^{-4} | 0.022 |

TABLE III. Performance of Newton’s, quasi-Newton and the fixed-point algorithm to solve the reduced system of equations defining the BiCM, on a set of real-world BiBUNs (of which basic statistics as the total number of nodes, \(N\), the total number of links, \(L\), and the connectance, \(c = L/(N \cdot M)\), are provided). All algorithms stop because the condition \(||\nabla \mathcal{L}(\hat{\theta})||_2 \leq 10^{-10}\) is satisfied. For what concerns both accuracy and speed, the best performing method is Newton’s one, followed by the quasi-Newton and the fixed-point recipes. Only the results corresponding to the best choice of initial conditions are reported.

For what concerns accuracy, the largest maximum error per method spans an interval (across all configurations) that amounts at \(\text{MADE}_{\text{Newton}} \simeq 10^{-13}\), \(\text{MADE}_{\text{Quasi-Newton}} \simeq 10^{-5}\) and \(\text{MADE}_{\text{Fixed-point}} \simeq 10^{-4}\). By looking at each specific network, it is evident that the most accurate method is systematically Newton’s one.

For what concerns speed, the amount of time required by each method to achieve convergence spans an interval (across all configurations) that is \(T_{\text{Newton}} \simeq 0.0023\) (on average), \(T_{\text{Quasi-Newton}} \simeq 0.016\) (on average) and \(T_{\text{Fixed-point}} \simeq 0.018\) (on average) (time is measured in seconds). The fastest method is Newton’s one and is followed by the quasi-Newton and the fixed-point recipes. The gain in terms of speed due to the reducibility (now quantified by the ratio between the number of different pairs of degrees divided by the total number of nodes) of the system of equations defining the BiCM is also evident: while solving the original problem would have required handling a system of \(\sim 10^3\) equations, the reduced one is defined by just \(\sim 10^2\) distinct equations. Overall, a solution is always found within thousandths or hundredths of seconds.

As for the DBCM case, the ensemble of BiBUNs can be sampled by implementing a Bernoulli trial \(p_{i\alpha} \sim \text{Ber}[p_{i\alpha}]\) for any two nodes (belonging to different layers) in either a sequential or a parallel fashion. The computational complexity of the sampling process amounts at \(O(N \cdot M)\) and can be repeated to generate as many configurations as desired. The pseudo-code for explicitly sampling the
BiCM ensemble is summed up by Algorithm 3.

**Algorithm 3 Sampling the BiCM ensemble**

1: for m = 1...|E| do
2: \( B = 0 \);
3: for \( i = 1...N \) do
4: \( \text{for } a = 1...M \) do
5: if RandomUniform[0,1] \( \leq p_{ia}^{\text{BiCM}} \) then
6: \( b_{ia} = 1; \)
7: else
8: \( b_{ia} = 0; \)
9: end if
10: end for
11: end for
12: Ensemble[m] = B;
13: end for

UECM: weighted undirected graphs with given strengths and degrees

So far, we have considered purely binary models. Let us now focus on a class of ‘mixed’ null models for weighted networks, defined by constraining both binary and weighted quantities. Purely weighted models such as the Undirected and the Directed Weighted Configuration Model have not been considered since, as it has been proven elsewhere [35], they perform quite poorly when employed to reconstruct networks. Let us start by the simplest model, i.e. the one constraining the degrees and the strengths in an undirected fashion. While \( k_i(A) = \sum_{j \neq i} a_{ij} \) counts the number of neighbors of node \( i \), \( s_i(W) = \sum_{j \neq i} w_{ij} \) defines the weighted equivalent of the degree of node \( i \), i.e. its strength. For consistency, the binary adjacency matrix can be defined via the Heaviside step function, \( \Theta[\cdot] \), as \( A \equiv \Theta[W] \) a position indicating that \( a_{ij} = 1 \) if \( w_{ij} > 0 \), \( \forall \ i < j \) and zero otherwise. This particular model is known as Undirected Enhanced Configuration Model (UECM) [35][40] and its Hamiltonian reads

\[
\mathcal{H}_{\text{UECM}}(W, \bar{\alpha}, \bar{\beta}) = \sum_{i=1}^{N} [\alpha_i k_i(A) + \beta_i s_i(W)];
\]

it induces a probability distribution which is halfway between a Bernoulli and a geometric one [39], i.e.

\[
Q_{\text{UECM}}(W|\bar{\alpha}, \bar{\beta}) = \prod_{i=1}^{N} \prod_{j=1}^{N} q_{ij}(w_{ij})
\]

with

\[
q_{ij}(w) = \begin{cases} 
1 - p_{ij}^{\text{UECM}}, & w = 0 \\
p_{ij}^{\text{UECM}}(e^{-\beta_i - \beta_j})^{w-1}(1 - e^{-\beta_i - \beta_j}), & w > 0
\end{cases}
\]

for any two nodes \( i \) and \( j \) such that \( i < j \) and \( p_{ij}^{\text{UECM}} = \frac{e^{-\alpha_i - \alpha_j - \beta_i - \beta_j}}{1 - e^{-\alpha_i - \alpha_j - \beta_i - \beta_j}} \). Notice that the functional form above is obtained upon requiring that the weights only assume (non-negative) integer values (i.e. \( w_{ij} \in [0, +\infty) \), \( \forall \ i < j \)): hence, the canonical ensemble is now constituted by the weighted configurations with \( N \) nodes and a number of (undirected) links ranging between zero and the maximum value \( \binom{N}{2} \).

The argument of the problem (14) for the specific network \( W^* \) now becomes

\[
\mathcal{L}_{\text{UECM}}(\bar{\alpha}, \bar{\beta}) = -\sum_{i=1}^{N} [\alpha_i k_i(A^*) + \beta_i s_i(W^*)] - \sum_{i=1}^{N} \sum_{j=1}^{N} \ln \left[ 1 + e^{-\alpha_i - \alpha_j} \frac{e^{-\beta_i - \beta_j}}{1 - e^{-\beta_i - \beta_j}} \right]
\]

whose first-order optimality conditions read

\[
\nabla_{\alpha_i} \mathcal{L}_{\text{UECM}} = -k_i(A^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{UECM}}
\]

\[
= -k_i(A^*) + \langle k_i \rangle = 0, \quad i = 1...N,
\]

\[
\nabla_{\beta_i} \mathcal{L}_{\text{UECM}} = -s_i(W^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{UECM}}
\]

\[
= -s_i(W^*) + \langle s_i \rangle = 0, \quad i = 1...N.
\]

**Resolution of the UECM.** Newton’s and the quasi-Newton methods can be easily implemented via the recipe defined in eq. (18) (see Appendix A for the definition of the UECM Hessian).

As for the purely binary models, the fixed-point recipe for solving the UECM first-order optimality conditions transforms the following set of consistency equations

\[
\alpha_i = -\ln \left[ \frac{k_i(A^*)}{\sum_{j \neq i}^{N} \frac{e^{-\alpha_j - \alpha_i - \beta_j}}{1 - e^{-\alpha_j - \alpha_i - \beta_j}}} \right],
\]

\[
\beta_i = -\ln \left[ \frac{s_i(W^*)}{\sum_{j \neq i}^{N} \frac{e^{-\alpha_i - \alpha_j - \beta_i}}{1 - e^{-\alpha_i - \alpha_j - \beta_i}}} \right]
\]

(with \( i = 1...N \)) into the usual iterative fashion, by considering the parameters at the left hand side and at the right hand side, respectively at the \( n \)-th and at the \( (n-1) \)-th iteration. It is important to remark that a
Table IV. Performance of Newton’s and the quasi-Newton method to solve the reduced system of equations defining the UECM, on a set of real-world WUNs (of which basic statistics as the total number of nodes, $N$, the total number of links, $L$, and the connectance, $c = 2L/N(N-1)$, are provided). While Newton’s method stops because the condition $\|\nabla \mathcal{L}(\hat{\theta})\|_2 \leq 10^{-8}$ is satisfied, the quasi-Newton one always reaches the limit of 10000 steps. The results on accuracy and speed clearly indicate that Newton’s method outperforms the quasi-Newton one. Only the results corresponding to the best choice of initial conditions are reported. The results of the fixed-point recipe are not shown.

![Table IV](https://example.com/table-iv.png)

The accuracy of each algorithm in reproducing the constraints defining the UECM has been now quantified via the maximum relative error, metrics, defined, in a perfectly general fashion, as $\max_{i} \left\{ \frac{C_i - C_i^*}{C_i} \right\}^N$, (where $C_i^*$ is the empirical value of the $i$-th constraint, $C_i$). In the UECM case, we can define two variants of the aforementioned error, i.e.

\[
\text{MRDE} = \max_{i} \left\{ \frac{|k_i^* - \langle k_i \rangle|}{k_i} \right\}_{i=1}^{N} \quad \text{(57)}
\]

\[
\text{MRSE} = \max_{i} \left\{ \frac{|s_i^* - \langle s_i \rangle|}{s_i} \right\}_{i=1}^{N} \quad \text{(58)}
\]

The results about the performance of our three algorithms are reported in Table IV. Overall, two out of three algorithms (i.e. Newton’s and the quasi-Newton methods) perform very satisfactorily, being accurate, fast and scalable; the third one (i.e. the fixed-point recipe), instead, performs very poorly. Moreover, while Newton’s
method stops because the condition on the norm of the likelihood is satisfied, both the quasi-Newton and the fixed-point algorithms are always found to satisfy the limit condition on the number of steps (i.e. they run for 10000 steps and, then, stop).

For what concerns accuracy, the largest maximum error made by Newton’s method (across all configurations) amounts at $10^{-10} \lesssim \text{MRDE}_{\text{Newton}} \lesssim 10^{-5}$ and $\text{MRSE}_{\text{Newton}} \gtrsim 10^{-10}$; on the other hand, the largest maximum error made by the quasi-Newton method (across all configurations) amounts at $10^{-5} \lesssim \text{MRDE}_{\text{Quasi-Newton}} \lesssim 10^{-1}$ and $10^{-5} \lesssim \text{MRSE}_{\text{Quasi-Newton}} \lesssim 10^{-4}$. For what concerns speed, Newton’s method employs tenths of seconds to achieve convergence on each configuration while the quasi-Newton one always requires tens of seconds (specifically, almost thirty seconds for each considered configuration). The results above indicate that the fastest and two most accurate method is systematically Newton’s one, suggesting that the ‘complexity’ of the model is such that the information encoded into the Hessian matrix cannot be ignored without consequences on the quality of the solution. The fixed-point algorithm, instead, stops after seconds but is affected by errors whose order of systematically magnitude amounts at $\text{MRDE}_{\text{fixed-point}} \approx 10^2$ and $1 \lesssim \text{MRSE}_{\text{fixed-point}} \lesssim 10^2$.

We also explicitly notice that the MADE basically coincides with the MRDE for all considered configurations, meaning that the largest error, made by the algorithms considered here to solve the UECM, affects the nodes with the lowest degree (i.e. equal to one). On the other hand, strengths are affected by a larger absolute error (i.e. the MASE, defined as $\text{MASE} = \max_i \{ |s_i^* - \langle s_i \rangle_1 \} \}$) than the degrees: if we calculate the MRSE, however, we realize that the largest errors affect very large strengths - hence being perfectly acceptable. For example, let us consider the WTW in 1993: the MASE amounts at 0.1 but, as the MRSE reveals, it affects a strength of the order of $10^3$.

Lastly, differences in the speed of convergence of the two methods discussed in this section, caused by the choice of a particular set of initial conditions, are observable: the ‘uniform’ prescription outperforms the other ones.

Finally, let us comment on the algorithm to sample the UECM ensemble and that can be compactly achieved by implementing a two-step procedure. Let us look back at the formal expression for the pair-specific probability distribution characterizing the UECM: it induces coefficients reading

$$
\begin{cases}
1 - p_{ij}^\text{UECM}, & w = 0 \\
p_{ij}^\text{UECM}(1 - e^{-\beta_i - \beta_j}), & w = 1 \\
p_{ij}^\text{UECM}(e^{-\beta_i - \beta_j})(1 - e^{-\beta_i - \beta_j}), & w = 2 \\
p_{ij}^\text{UECM}(e^{-\beta_i - \beta_j})^2(1 - e^{-\beta_i - \beta_j}), & w = 3 \\
& \vdots
\end{cases}
$$

in turn suggesting that, for a specific pair of vertices $i, j$ (with $i < j$), the appearance of the first link is ruled by a Bernoulli distribution with probability $p_{ij}^\text{UECM}$ while the remaining $(w - 1)$ ones can be drawn from a geometric distribution whose parameter reads $e^{-\beta_i - \beta_j}$; in other words, the weight $(w - 1)$ is drawn conditionally on the presence of a connection between the two considered nodes. The computational complexity of the sampling process is, again, $O(N^2)$. The pseudo-code for explicitly sampling the DBCM ensemble is summed up by Algorithm 4. Notice that the way our sampling procedure is written requires the support of the geometric distribution to coincide with the positive integers.

Algorithm 4 Sampling the UECM ensemble

1. for $m = 1 \ldots |E|$ do
2. $W = 0$;
3. for $i = 1 \ldots N$ do
4. for $j = 1 \ldots N$ and $j < i$ do
5. if RandomUniform$[0, 1] \leq p_{ij}^\text{UECM}$ then
6. $w_{ij} = w_{ji} = \text{RandomGeometric}[e^{-\beta_i - \beta_j}]$;
7. else
8. $w_{ij} = w_{ji} = 0$;
9. end if
10. end for
11. end for
12. Ensemble$[m] = W$;
13. end for

DECM: weighted directed graphs with given strengths and degrees

Let us now extend the ‘mixed’ model introduced in the previous section to the case of directed networks. Constraints are, now, represented by four sequences of values, i.e. $(k_{out}^i)_i \in [1, N], (k_{out}^{in})_i \in [1, N], (s_{out}^i)_i \in [1, N], (s_{in}^{out})_i \in [1, N]$ where the generic out-degree and in-degree are, respectively, defined as $k_{out}^i(A) = \sum_{j \neq i=1}^N a_{ij}$ and $k_{out}^{in}(A) = \sum_{j \neq i=1}^N a_{ji}$ and analogously for the generic out-strength and in-strength, reading $s_{out}^i(W) = \sum_{j \neq i=1}^N w_{ij}$ and $s_{in}^{out}(W) = \sum_{j \neq i=1}^N w_{ji}$. Consistency requires that $A \equiv \Theta[W]$ as for the UECM case. This model is known as Directed Enhanced Configuration Model (DECM) and its Hamiltonian reads
\[ \mathcal{H}_{\text{DECM}}(\mathbf{W}, \alpha, \beta, \gamma, \delta) = \sum_{i=1}^{N} (\alpha_i k_i^\text{out}(\mathbf{A}) + \beta_i k_i^\text{in}(\mathbf{A})) + \gamma_i s_i^\text{out}(\mathbf{W}) + \delta_i s_i^\text{in}(\mathbf{W}) \] (60)

in turn, inducing the directed counterpart of the UECM distribution, i.e.

\[ Q_{\text{DECM}}(\mathbf{W}|\alpha, \beta, \gamma, \delta) = \prod_{i=1}^{N} \prod_{j=1}^{N} q_{ij}(w_{ij}) \] (61)

with

\[ q_{ij}(w) = \begin{cases} 1 - p_{ij}^{\text{DECM}} & w = 0 \\ p_{ij}^{\text{DECM}}(e^{-\gamma_i - \delta_j} - 1) & w > 0 \end{cases} \] (62)

for any two nodes \( i \) and \( j \) such that \( i \neq j \) and \( p_{ij}^{\text{DECM}} = \frac{e^{-\gamma_i - \delta_j} - 1}{e^{-\gamma_i} + e^{-\delta_j} - 1} \). As for the undirected case, weights are required to assume only (non-negative) integer values (i.e. \( w_{ij} \in [0, +\infty), \forall i \neq j \)): hence, the canonical ensemble is constituted by the weighted configurations with \( N \) nodes and a number of (directed) links ranging between zero and the maximum value \( N(N-1) \).

The argument of the problem (14) for the specific network \( \mathbf{W}^* \) becomes

\[ \mathcal{L}_{\text{DECM}}(\alpha, \beta, \gamma, \delta) = -\sum_{i=1}^{N} \left[ \alpha_i k_i^\text{out}(\mathbf{A}^*) + \beta_i k_i^\text{in}(\mathbf{A}^*) \right] \\
+ \gamma_i s_i^\text{out}(\mathbf{W}^*) + \delta_i s_i^\text{in}(\mathbf{W}^*) \] (63)

with \( z_{ij} = \left[ 1 + e^{-\alpha_i - \beta_j} \left( \frac{e^{-\gamma_i - \delta_j}}{1 - e^{-\gamma_i - \delta_j}} \right) \right], \forall i \neq j \) and whose first-order optimality conditions read

\[ \mathcal{L}_{\alpha, \mathcal{L}_{\text{DECM}}} = -k_i^\text{out}(\mathbf{A}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (64)

\[ \mathcal{L}_{\beta, \mathcal{L}_{\text{DECM}}} = -k_i^\text{in}(\mathbf{A}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (65)

\[ \mathcal{L}_{\gamma, \mathcal{L}_{\text{DECM}}} = -s_i^\text{out}(\mathbf{W}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (66)

\[ \mathcal{L}_{\delta, \mathcal{L}_{\text{DECM}}} = -s_i^\text{in}(\mathbf{W}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (67)

\[ \nabla_{\alpha_i} \mathcal{L}_{\text{DECM}} = -k_i^\text{out}(\mathbf{A}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (68)

\[ \nabla_{\beta_i} \mathcal{L}_{\text{DECM}} = -k_i^\text{in}(\mathbf{A}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (69)

\[ \nabla_{\gamma_i} \mathcal{L}_{\text{DECM}} = -s_i^\text{out}(\mathbf{W}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (70)

\[ \nabla_{\delta_i} \mathcal{L}_{\text{DECM}} = -s_i^\text{in}(\mathbf{W}^*) + \sum_{j \neq i}^{N} p_{ij}^{\text{DECM}} \] (71)

Resolution of the DECM. Newton’s and the quasi-Newton methods can be easily implemented via the recipe defined in eq. (18) (see Appendix A for the definition of the DECM Hessian).

As for the UECM, the fixed-point recipe for solving the DECM first-order optimality conditions transforms the following set of consistency equations

\[ \alpha_i = -\ln \left[ \frac{k_i^\text{out}(\mathbf{A}^*)}{\sum_{j \neq i}^{N} \left( \frac{e^{-\beta_j - \gamma_i - \delta_j}}{1 - e^{-\gamma_i - \delta_j}} \right)} \right] , \] (72)

\[ \beta_i = -\ln \left[ \frac{k_i^\text{in}(\mathbf{A}^*)}{\sum_{j \neq i}^{N} \left( \frac{e^{-\alpha_j - \gamma_i - \delta_i}}{1 - e^{-\gamma_i - \delta_i}} \right)} \right] , \] (73)

\[ \gamma_i = -\ln \left[ \frac{s_i^\text{out}(\mathbf{W}^*)}{\sum_{j \neq i}^{N} \left( \frac{e^{-\alpha_j - \beta_i}}{1 - e^{-\gamma_i - \delta_j}} \right)} \right] , \] (74)

\[ \delta_i = -\ln \left[ \frac{s_i^\text{in}(\mathbf{W}^*)}{\sum_{j \neq i}^{N} \left( \frac{e^{-\alpha_j - \beta_i}}{1 - e^{-\gamma_i - \delta_j}} \right)} \right] . \] (75)

(65)
than the purely binary models we have considered in the present contribution.

The three different sets of initial conditions that have been considered generalize the UECM ones: in particular, the first set of values reads

\[ \alpha_i^{(0)} = - \ln \left[ \frac{k_{i-1}^{(\alpha)}/\sqrt{L}}{k_i} \right], \quad i = 1 \ldots N, \quad \beta_i^{(0)} = - \ln \left[ \frac{\gamma_{i-1}^{(\beta)}}{\sqrt{N}} \right], \quad i = 1 \ldots N, \]

\[ \gamma_i^{(0)} = - \ln \left[ \frac{\delta_{i-1}^{(\gamma)}}{\sqrt{W}} \right], \quad i = 1 \ldots N; \quad \delta_i^{(0)} = - \ln \left[ \frac{\delta_{i-1}^{(\delta)}}{\sqrt{W}} \right], \quad i = 1 \ldots N; \]

The second set of initial conditions can be obtained by simply replacing \( L \) with \( N \); the third recipe, as usual, prescribes to randomly draw the value of each parameter from the uniform distribution defined on the unit interval.

**Performance testing.** The performance of the three algorithms to solve the system of equations defining the DECM has been tested on a bunch of real-world networks. In particular, we have considered the Electronic Italian Interbank Market (e-MID) during the decade 2000-2010 [42]. Since e-MID weights are real numbers, we have rounded them to the nearest integer value, before running the DECM. Before commenting on the results of our numerical exercises, let us, first, describe how the latter ones have been carried out.

The accuracy of each algorithm in reproducing the constraints defining the DECM has been quantified via the maximum relative error metrics, now reading

\[
\text{MRDE} = \max_i \left\{ \left| k_i^{*} - \langle k_i \rangle \right| / k_i \right\}^{N} \quad \text{and} \quad \text{MRSE} = \max_i \left\{ \left| s_i^{*} - \langle s_i \rangle \right| / s_i \right\}^{N} \tag{66} \tag{67}
\]

The results about the performance of our three algorithms are reported in Table 5. Overall, Newton’s method performs very satisfactorily, being accurate, fast and scalable; the quasi-Newton method is accurate as well although (in some cases, much) slower. The fixed-point recipe, instead, performs very poorly, as for the undirected case. Moreover, while Newton’s method stops because the condition on the norm of the likelihood is satisfied, both the quasi-Newton and the fixed-point algorithms are always found to satisfy the limit condition on the number of steps (i.e. after 10000 steps, any of the three algorithms stops).

The three different ‘stop criteria’ we have adopted are the same ones we have considered for both the binary and the undirected, ‘mixed’ model, i.e. the condition on the Euclidean norm of the gradient of the likelihood function, i.e. \( ||\mathcal{L}(\hat{\theta})||_2 \leq 10^{-8} \), the condition on the gradient of the likelihood function, i.e. \( ||\Delta \theta||_2 \leq 10^{-8} \) and the condition on the maximum number of iterations (i.e. after 10000 steps, any of the three algorithms stops).

For what concerns accuracy, the largest maximum error made by Newton’s method (across all configurations) amounts to \( 10^{-14} \lesssim \text{MRDE}_{\text{Newton}} \lesssim 10^{-7} \) and \( 10^{-12} \lesssim \text{MRSE}_{\text{Newton}} \lesssim 10^{-5} \); on the other hand, the largest maximum error made by the quasi-Newton method (across all configurations) amounts to \( 10^{-8} \lesssim \text{MRDE}_{\text{Quasi-Newton}} \lesssim 10^{-7} \) and \( 10^{-4} \lesssim \text{MRSE}_{\text{Quasi-Newton}} \lesssim 10^{-3} \). For what concerns speed, Newton’s method employs tens of seconds to achieve convergence on each configuration; the

| N  | L     | c       | MRDE | MASE | MRSE | Time (s) | MRDE | MASE | MRSE | Time (s) |
|----|-------|---------|------|------|------|----------|------|------|------|----------|
| e-MID 00 | 196 | 10618  | 0.28 | 1.4·10^{-5} | 5·10^{-8} | 1.5·10^{-10} | 0.5  | 1.7·10^{-11} | 3·10^{-12} | 5.4·10^{-14} | 0.1 |
| e-MID 01 | 185 | 8951   | 0.26 | 1.4·10^{-11} | 6·10^{-9} | 2·10^{-10} | 0.4  | 1.4·10^{-11} | 3·10^{-12} | 7·10^{-14} | 0.2 |
| e-MID 02 | 177 | 7252   | 0.23 | 1.4·10^{-9} | 10^{-8} | 1·10^{-10} | 0.5  | 9.5·10^{-8} | 6·10^{-10} | 7.4·10^{-12} | 0.1 |
| e-MID 03 | 179 | 6814   | 0.21 | 1.6·10^{-9} | 2·10^{-8} | 4.4·10^{-10} | 0.9  | 9.6·10^{-8} | 50   | 1.1·10^{-8} | 0.2 |
| e-MID 04 | 180 | 6136   | 0.19 | 6.5·10^{-6} | 9·10^{-10} | 3.4·10^{-12} | 0.9  | 1·10^{-11} | 700  | 4.2·10^{-13} | 0.2 |
| e-MID 05 | 176 | 6203   | 0.21 | 3·10^{-14} | 1·10^{-9} | 9.4·10^{-11} | 1.2  | 4.8·10^{-8} | 300  | 2.4·10^{-8} | 0.3 |
| e-MID 06 | 177 | 6132   | 0.19 | 1.5·10^{-14} | 2·10^{-10} | 1.8·10^{-12} | 0.9  | 5·10^{-10} | 60   | 2.5·10^{-10} | 0.2 |
| e-MID 07 | 178 | 6330   | 0.21 | 8.4·10^{-7} | 1·10^{-13} | 2.5·10^{-16} | 0.7  | 1.8·10^{-8} | 3   | 7.6·10^{-9} | 0.1 |
| e-MID 08 | 173 | 4767   | 0.16 | 2·10^{-9} | 2·10^{-9} | 7.7·10^{-13} | 0.6  | 1.8·10^{-8} | 20  | 1.2·10^{-8} | 0.3 |
| e-MID 09 | 156 | 2961   | 0.12 | 2.6·10^{-13} | 1·10^{-13} | 4.9·10^{-16} | 0.6  | 1.7·10^{-8} | 1   | 9·10^{-10} | 0.1 |
| e-MID 10 | 135 | 2743   | 0.15 | 6.3·10^{-8} | 3·10^{-8} | 5.9·10^{-10} | 0.7  | 1.6·10^{-8} | 4   | 5.2·10^{-9} | 0.1 |

** TABLE V. ** Performance of Newton’s and the quasi-Newton method to solve the reduced system of equations defining the DECM, on a set of real-world WDNs (of which basic statistics as the total number of nodes, \( N \), the total number of links, \( L \), and the connectance, \( c \), \( = L/N(N-1) \), are provided). While Newton’s method stops because the condition \( ||\mathcal{L}(\hat{\theta})||_2 \leq 10^{-8} \) is satisfied, the quasi-Newton one always reaches the limit of 10000 steps. The results on accuracy and speed clearly indicate that Newton’s method outperforms the quasi-Newton one. Only the results corresponding to the best choice of initial conditions are reported. The results of the fixed-point recipe are not shown.
time required by the quasi-Newton method is of the same order of magnitude, although it is systematically larger than the time required by Newton’s one. Overall, these results indicate that the fastest and two most accurate method is Newton’s one. As in the undirected case, the fixed-point algorithm, instead, stops after seconds but is affected by errors whose order of systematically magnitude amounts at $10 \lesssim \text{MRDE}_{\text{fixed-point}} \lesssim 10^2$ and $1 \lesssim \text{MRSE}_{\text{fixed-point}} \lesssim 10^2$.

As for the UECM, the MADE basically coincides with the MRDE, for all considered configurations, while strengths are affected by a larger absolute error than the MRSE, we realize that the largest errors affect very large strengths - hence being perfectly acceptable.

Lastly, differences in the speed of convergence of the two methods discussed in this section, caused by the choice of a particular set of initial conditions, are observable: the ‘uniform’ prescription outperforms the other ones.

Finally, let us comment on the algorithm to sample the DECM ensemble: as for the UECM, it can be compactly achieved by implementing the directed counterpart of the two-step procedure described above. Given a specific pair of vertices $i, j$ (with $i \neq j$), the first link can be drawn by sampling a Bernoulli distribution with probability $p_{ij}^{\text{DEC}}$ while the remaining $(w - 1)$ ones can be drawn from a geometric distribution whose parameter reads $e^{-\gamma_i - \delta_j}$.

The computational complexity of the sampling process is, again, $O(N^2)$ and the pseudo-code for explicitly sampling the DECM ensemble is summed up by Algorithm 5. Notice that the way our sampling procedure is written requires the support of the geometric distribution to coincide with the positive integers.

**Algorithm 5 Sampling the DECM ensemble**

1: for $m = 1 \ldots |E|$ do
2: $W = 0$;
3: for $i = 1 \ldots N$ do
4: for $j = 1 \ldots N$ and $j \neq i$ do
5: if $\text{RandomUniform}[0, 1] \leq p_{ij}^{\text{DEC}}$ then
6: \hspace{1em} $w_{ij} = \text{RandomGeometric}[e^{-\gamma_i - \delta_j}]$;
7: else
8: \hspace{1em} $w_{ij} = 0$;
9: end if
10: end for
11: end for
12: Ensemble[$m$] = $W$;
13: end for

**Two-step models for undirected and directed networks**

The need of considering network models defined in a two-step fashion arises from a number of considerations. First, the amount of information concerning binary and weighted quantities is often asymmetric: as it has been pointed out in [43], information concerning a given network structure ranges from the knowledge of just a single, aggregated piece of information (e.g. the link density) to that of entire subgraphs. Indeed, models exist that take as input any binary, either probabilistic or deterministic, network model (i.e. any $P(A)$) while placing link weights optimally, conditionally on the input configurations [10][43].

Second, recipes like the UECM and the DECM are, generally speaking, difficult to solve; as we have already observed, only Newton’s method performs in a satisfactory way, both for what concerns accuracy and speed: hence, easier-to-solve recipes are welcome.

In what follows, we will consider the conditional reconstruction method (hereby, CReM) induced by the Hamiltonian

$$H_{\text{CReM}}(W, \theta) = \sum_{i=1}^{N} \theta_i s_i(A);$$

in case of undirected networks, it induces a conditional probability distribution reading

$$Q(W|A) = \prod_{i=1}^{N} \prod_{j=1}^{N} q_{ij}(w_{ij}|a_{ij})$$

where, for consistency, $q_{ij}(w_{ij} = 0|a_{ij} = 0) = 1$ and $q_{ij}(w_{ij} = 0|a_{ij} = 1) = 0$. The meaning of these relationships is the following: given any two nodes $i$ and $j$, the absence of a link, i.e. $a_{ij} = 0$, admits the only possibility $w_{ij} = 0$; on the other hand, the presence of a link, i.e. $a_{ij} = 1$, rules out the possibility that a null weight among the same vertices is observed.

In general, the functional form of $q_{ij}(w_{ij}|a_{ij} = 1)$ depends on the domain of the weights. In all cases considered in [10][43], weights are assumed to be continuous; since the continuous distribution that maximizes Shannon entropy while constrained to reproduce first-order moments is the exponential one, the following functional form

$$q_{ij}(w_{ij}|a_{ij} = 1) = \begin{cases} (\theta_i + \theta_j) e^{-(\theta_i + \theta_j)w} & w > 0 \\ 0 & w \leq 0 \end{cases}$$

(for any undirected pair of nodes) remains naturally induced. As shown in [43], the problem [14] has to be slightly generalized; still, its argument for the specific network $W^*$ becomes

$$G_{\text{CReM}} = -\sum_{i=1}^{N} \theta_i s_i(W^*) + \sum_{i=1}^{N} \sum_{j=1}^{N} f_{ij} \log[\theta_i + \theta_j]$$
The first set of initial values reads \( \theta_i^{(0)} = -\ln \left[ \frac{s_i(W^*)}{\sqrt{2 \pi W}} \right], \) \( i = 1 \ldots N; \) the second one is a variant of the position constraint above, reading \( \theta_i^{(0)} = -\ln \left[ \frac{s_i(W^*)}{\sqrt{N}} \right]; \) the third one, instead, prescribes to randomly draw the value of each parameter from the uniform distribution defined on the unit interval, i.e. \( \theta_i^{(0)} \sim U(0, 1), \forall i. \)

When considering directed networks, the conditional probability distribution defining the CRm reads

\[
q_{ij}(w_{ij}|a_{ij} = 1) = \begin{cases} 
(a_i + \beta_j)e^{-(a_i + \beta_j)w} & w > 0 \\
0 & w \leq 0
\end{cases}
\]

for any two nodes \( i \) and \( j \) such that \( i \neq j; \) the set of equations (73) can be generalized as follows

\[
\alpha_i = \left[ \frac{\sum_{j=1}^{N} s_{ij}^{\text{out}}(W^*)}{\sum_{j=1}^{N} \frac{f_{ij}}{1 + \beta_j/\alpha_j}} \right]^{-1}, \quad i = 1 \ldots N
\]

\[
\beta_i = \left[ \frac{\sum_{j=1}^{N} s_{ij}^{\text{in}}(W^*)}{\sum_{j=1}^{N} \frac{f_{ij}}{1 + \alpha_i/\beta_j}} \right]^{-1}, \quad i = 1 \ldots N
\]

and analogously for the sets of values initializing them.

**Rescaling the CRm algorithm.** Although the equations defining the CRm algorithm cannot be effectively reduced, they can be opportunistically rescaled. To this aim, let us consider directed configurations and the system

\[
\begin{align*}
\sum_{j=1}^{N} \frac{f_{ij}}{\alpha_i(\kappa) + \beta_j(\kappa)} &= \frac{s_{ij}^{\text{out}}(W^*)}{\kappa}, \quad \forall i \\
\sum_{j=1}^{N} \frac{f_{ij}}{\alpha_i(\kappa) + \beta_j(\kappa)} &= \frac{s_{ij}^{\text{in}}(W^*)}{\kappa}, \quad \forall i
\end{align*}
\]

where the quantity \( f_{ij} = \sum_A P(A)a_{ij} \) represents the expected value of \( a_{ij} \) over the ensemble of binary configurations realizing the binary model taken as input (i.e. the marginal probability of an edge existing between nodes \( i \) and \( j \)). It follows that the CRm first-order optimality conditions read

\[
\nabla_{\theta_i} \mathcal{L}_{\text{CRm}} = -s_i(W^*) + \sum_{j\neq i=1}^{N} \frac{f_{ij}}{\theta_i + \theta_j} = -s_i(W^*) + \langle s_i \rangle = 0, \quad i = 1 \ldots N.
\]

**Resolution of the CRm.** Newton’s and the quasi-Newton method can still be implemented via the recipe defined in eq. (18) (see Appendix A for the definition of the CRm Hessian).

As for the UECM and the DECM, the fixed-point recipe for solving the system of equations embodying the CRm transforms the set of consistency equations

\[
\theta_i = \left[ \frac{s_i(W^*)}{\sum_{j=1}^{N} \frac{f_{ij}}{1 + \beta_j/\alpha_i}} \right]^{-1}, \quad i = 1 \ldots N
\]

into an iterative recipe of the usual form, i.e. by considering the parameters at the left hand side and at the right hand side, respectively at the \( n \)-th and at the \( (n - 1) \)-th iteration. Although a reduced recipe can, in principle, be defined, an analogous observation to the one concerning the UECM and the DECM holds: the mathematical nature of the strengths (now, real numbers) increases their heterogeneity, in turn causing the CRm algorithm to be reducible even less than the ‘mixed’ models defined by discrete weights.

The initialization of the iterative recipe for solving the CRm has been implemented in the usual threefold way.
where the sufficient statistics has been divided by an opportune defined factor (in this case, $\kappa$) and the symbols $\alpha_i(\kappa)$, $\beta_j(\kappa)$, $\beta_j(\kappa)$ and $\beta_j(\kappa)$ that the solution we are searching for is a function of the parameter $\kappa$ itself. In fact, a solution of the system above reads

$$
\alpha_i^* = \frac{\alpha_i(\kappa)}{\kappa}, \quad \forall i \tag{77}
$$

$$
\beta_i^* = \frac{\beta_i(\kappa)}{\kappa}, \quad \forall i \tag{78}
$$

as it can be proven upon substituting it back into eqs. (76) and noticing that $\{\alpha_i^*\}_{i=1}^N$ and $\{\beta_i^*\}_{i=1}^N$ are solutions of the system of equations defined in (76). As our likelihood maximization problem admits a unique, global maximum, the prescription above allows us to easily identify it. Rescaling will be tested in order to find out if our algorithms are enhanced by it under some respect (e.g. accuracy or speed).

**Performance testing.** Before commenting on the performance of the three algorithms in solving the system of equations defining the CREM, let us stress once more that the formulas presented so far are perfectly general, working for any binary recipe one may want to employ. In what follows, we will test the CREM by posing $f_{ij} \equiv p_{ij}^{UBCM}$ and $f_{ij} \equiv p_{ij}^{DBC M}$.

To test the effectiveness of our algorithms in solving the CREM on undirected networks we have considered the synaptic network of the worm *C. Elegans* [28] and the eight daily snapshots of the Bitcoin Lightning Network (e-MID) during the decade 2000-2010 [42]. Before commenting on the results of our numerical exercises, let us, first, describe how the latter ones have been carried out.

As for the discrete ‘mixed’ models, the accuracy of each algorithm in reproducing the constraints defining the CREM has been quantified via the Maximum Relative Degree Error and the Maximum Relative Strength Error metrics, whose definition is provided by eqs. (57), (58) and (60), (67) for the undirected and the directed case, respectively. Analogously, the three ‘stop criteria’ for each algorithm are the same ones that we have adopted for the other models (and consist in a condition on the Euclidean norm of the gradient of the likelihood function, i.e. $||\nabla \mathcal{L}(\theta)||_2 \leq 10^{-8}$, a condition on the Euclidean norm of the vector of differences between the values of the parameters at subsequent iterations, i.e. $||\Delta \theta||_2 \leq 10^{-8}$, and a condition on the maximum number of iterations, i.e. 10000 steps).

The results about the performance of our three algorithms are reported in Table VI and in Table VII. Let us start by commenting the results reported in Table VI and concerning undirected networks. Generally speaking, Newton’s method is the most accurate one (its largest maximum errors span intervals, across all configurations, that amount at $10^{-11} \lesssim \text{MRDE}_{\text{Newton}} \lesssim 10^{-8}$ and $10^{-5} \lesssim \text{MRSE}_{\text{Newton}} \lesssim 10^{-4}$) although it scales very badly with the size of the network on which it is tested (the amount of time, measured in seconds, required by it to achieve convergence spans an interval, across all configurations, that amounts at $0.08 \leq T_{\text{Newton}} \leq 1188$).

The quasi-Newton method, on the other hand, is very accurate on the degrees (as already observed in the UBCM case) but not so accurate in reproducing the weighted constraints (its largest maximum errors span intervals, across all configurations, that amount at $\text{MRDE}_{\text{Quasi-Newton}} \simeq 10^{-7}$ and $10^{-6} \lesssim \text{MRSE}_{\text{Quasi-Newton}} \lesssim 6$). Moreover, it scales even worse

| $N$  | $L$  | MRDE | MRSE | Time (s) | MRDE | MRSE | Time (s) | MRDE | MRSE | Time (s) |
|------|------|------|------|----------|------|------|----------|------|------|----------|
| e-MID 00 | 196 | 10618 | $\simeq 3 \cdot 10^{-11}$ | $\simeq 2 \cdot 10^{-10}$ | $\simeq 0.9$ | $\simeq 5 \cdot 10^{-9}$ | $\simeq 2 \cdot 10^{-9}$ | $\simeq 0.9$ | $\simeq 4 \cdot 10^{-13}$ | $\simeq 8 \cdot 10^{-3}$ | $\simeq 0.09$ |
| e-MID 01 | 185 | 8051 | $\simeq 3 \cdot 10^{-11}$ | $\simeq 1 \cdot 10^{-10}$ | $\simeq 0.9$ | $\simeq 7 \cdot 10^{-10}$ | $\simeq 5 \cdot 10^{-10}$ | $\simeq 1$ | $\simeq 7 \cdot 10^{-13}$ | $\simeq 1 \cdot 10^{-4}$ | $\simeq 0.12$ |
| e-MID 02 | 177 | 7252 | $\simeq 6 \cdot 10^{-12}$ | $\simeq 2 \cdot 10^{-13}$ | $\simeq 0.7$ | $\simeq 3 \cdot 10^{-10}$ | $\simeq 5 \cdot 10^{-10}$ | $\simeq 1$ | $\simeq 8 \cdot 10^{-10}$ | $\simeq 3 \cdot 10^{-4}$ | $\simeq 0.08$ |
| e-MID 03 | 179 | 6814 | $\simeq 1 \cdot 10^{-12}$ | $\simeq 2 \cdot 10^{-10}$ | $\simeq 0.8$ | $\simeq 3 \cdot 10^{-10}$ | $\simeq 3 \cdot 10^{-10}$ | $\simeq 0.7$ | $\simeq 4 \cdot 10^{-10}$ | $\simeq 8 \cdot 10^{-4}$ | $\simeq 0.1$ |
| e-MID 04 | 180 | 6136 | $\simeq 9 \cdot 10^{-10}$ | $\simeq 2 \cdot 10^{-9}$ | $\simeq 0.9$ | $\simeq 5 \cdot 10^{-10}$ | $\simeq 8 \cdot 10^{-9}$ | $\simeq 0.8$ | $\simeq 5 \cdot 10^{-9}$ | $\simeq 6 \cdot 10^{-4}$ | $\simeq 0.13$ |
| e-MID 05 | 176 | 6203 | $\simeq 3 \cdot 10^{-10}$ | $\simeq 5 \cdot 10^{-9}$ | $\simeq 0.7$ | $\simeq 2 \cdot 10^{-10}$ | $\simeq 2 \cdot 10^{-10}$ | $\simeq 0.7$ | $\simeq 5 \cdot 10^{-9}$ | $\simeq 1 \cdot 10^{-4}$ | $\simeq 0.2$ |
| e-MID 06 | 177 | 6132 | $\simeq 1 \cdot 10^{-10}$ | $\simeq 7 \cdot 10^{-10}$ | $\simeq 0.7$ | $\simeq 2 \cdot 10^{-10}$ | $\simeq 3 \cdot 10^{-10}$ | $\simeq 0.8$ | $\simeq 8 \cdot 10^{-11}$ | $\simeq 5 \cdot 10^{-4}$ | $\simeq 0.14$ |
| e-MID 07 | 178 | 6330 | $\simeq 3 \cdot 10^{-10}$ | $\simeq 10^{-9}$ | $\simeq 1$ | $\simeq 1 \cdot 10^{-9}$ | $\simeq 8 \cdot 10^{-9}$ | $\simeq 1.2$ | $\simeq 7 \cdot 10^{-12}$ | $\simeq 7 \cdot 10^{-4}$ | $\simeq 0.14$ |
| e-MID 08 | 173 | 4767 | $\simeq 3 \cdot 10^{-10}$ | $\simeq 8 \cdot 10^{-10}$ | $\simeq 0.8$ | $\simeq 1 \cdot 10^{-9}$ | $\simeq 1 \cdot 10^{-9}$ | $\simeq 0.7$ | $\simeq 3 \cdot 10^{-10}$ | $\simeq 8 \cdot 10^{-4}$ | $\simeq 0.07$ |
| e-MID 09 | 156 | 2961 | $\simeq 4 \cdot 10^{-11}$ | $\simeq 3 \cdot 10^{-10}$ | $\simeq 0.6$ | $\simeq 1 \cdot 10^{-9}$ | $\simeq 9 \cdot 10^{-10}$ | $\simeq 0.7$ | $\simeq 8 \cdot 10^{-10}$ | $\simeq 2 \cdot 10^{-4}$ | $\simeq 0.11$ |
| e-MID 10 | 135 | 2743 | $\simeq 2 \cdot 10^{-10}$ | $\simeq 2 \cdot 10^{-9}$ | $\simeq 0.7$ | $\simeq 7 \cdot 10^{-10}$ | $\simeq 5 \cdot 10^{-10}$ | $\simeq 0.5$ | $\simeq 5 \cdot 10^{-10}$ | $\simeq 2 \cdot 10^{-10}$ | $\simeq 0.05$ |
than Newton’s method with the size of the network on which it is tested (the amount of time, measured in seconds, required by it to achieve convergence spans an interval, across all configurations, that amounts at $5 \leq T_{\text{Quasi-Newton}} \leq 15888$).

The performance of the fixed-point recipe is, somehow, intermediate between that of Newton’s and the quasi-Newton method. For what concerns accuracy, the fixed-point method is the least sensitive one to the growing size of the considered configurations: hence, it is also the fastest one (the amount of time, measured in seconds, required by it to achieve convergence spans an interval, across all configurations, that amounts at $0.01 \leq T_{\text{fixed-point}} \leq 550$).

Moreover, while Newton’s and the fixed-point method stop because the condition on the norm of the likelihood is satisfied, the quasi-Newton method is often found to satisfy the limit condition on the number of steps (i.e. it runs for 10000 steps and, then, stops).

Interestingly, the fact that the CReM cannot be reduced (at least not to a comparable extent with the one characterizing purely binary models) reveals a dependence on the network size of Newton’s and the quasi-Newton algorithms. The reason may lie in the evidence that both Newton’s and the quasi-Newton method require (some proxy of) the Hessian matrix of the system of equations defining the CReM to update the value of the parameters: as already observed, the order of the latter - which is $O(N^2)$ for Newton’s method and $O(N)$ for the quasi-Newton one - can make its calculation (very) time demanding.

Let us now move to comment on the performance of our algorithms when applied to solve the directed version of the CReM (see Table VII). Overall, all methods perform much better than in the undirected case, stopping because the condition on the norm of the likelihood is satisfied.

In fact, all of them are very accurate in reproducing the purely binary constraints, their largest maximum errors spanning intervals, across all configurations, that amount at $10^{-12} \lesssim \text{MRDE}_{\text{Newton}} \lesssim 10^{-6}$, $10^{-14} \lesssim \text{MRDE}_{\text{Quasi-Newton}} \lesssim 10^{-6}$ and $10^{-15} \lesssim \text{MRDE}_{\text{fixed-point}} \lesssim 10^{-8}$; for what concerns the weighted constraints, instead, the two most accurate methods are Newton’s and the quasi-Newton one, their largest maximum errors spanning intervals, across all configurations, that amount at $10^{-13} \lesssim \text{MRSE}_{\text{Newton}} \lesssim 10^{-7}$ and $10^{-6} \lesssim \text{MRSE}_{\text{Quasi-Newton}} \lesssim 10^{-3}$ (the fixed-point method performs worse than them, since $10^{-3} \lesssim \text{MRSE}_{\text{fixed-point}} \lesssim 10^{-1}$).

For what concerns speed, the amount of time, measured in seconds, required by Newton’s, the quasi-Newton and the fixed-point algorithms to achieve convergence spans an interval, across all configurations, that amounts at $0.6 \leq T_{\text{Newton}} \leq 1$, $0.5 \leq T_{\text{Quasi-Newton}} \leq 1.2$ and $0.05 \leq T_{\text{fixed-point}} \leq 0.2$, respectively. Hence, all methods are also very fast - the fixed-point being systematically the fastest one.

As already stressed above, the fact that the e-MID number of nodes remains approximately constant throughout the considered time interval masks the strong dependence of the performance of Newton’s and the quasi-Newton method on the network size.

Lastly, while rescaling the system of equations defining the CReM improves neither the accuracy nor the speed of any of the three algorithms considered here, differences in their speed of convergence, caused by the choice of a particular set of initial conditions, are observable: the ‘uniform’ prescription outperforms the other ones (for both the undirected and the directed version of the CReM).

As usual, let us comment on the algorithm to sample the CReM ensemble - for the sake of simplicity, on the undirected case. As for the UECM it can be compactly achieved by implementing a two-step procedure, the only difference lying in the functional form of the distribution from which weights are sampled. Given a specific pair of vertices $i, j$ (with $i < j$), the first link can be drawn from a Bernoulli distribution with probability $p_{ij}^{\text{UBCM}}$ while the remaining $(w-1)$ ones can be drawn from an exponential distribution whose parameter reads $\theta_i + \theta_j$. The computational complexity of the sampling process is, again, $O(N^2)$ and the pseudo-code for explicitly sampling the CReM ensemble is summed up by Algorithm 6.

**Algorithm 6 Sampling the CReM ensemble**

1: for $m = 1 \ldots |E|$ do
2: \hspace{1em} $W = 0$;
3: \hspace{1em} for $i = 1 \ldots N$ do
4: \hspace{2em} for $j = 1 \ldots N$ and $j < i$ do
5: \hspace{3em} if RandomUniform[0,1] $\leq p_{ij}^{\text{UBCM}}$ then
6: \hspace{4em} $w_{ij} = w_{ji} = \text{RandomExponential}[\theta_i + \theta_j]$;
7: \hspace{3em} else
8: \hspace{4em} $w_{ij} = w_{ji} = 0$;
9: \hspace{2em} end if
10: \hspace{1em} end for
11: \hspace{1em} end for
12: \hspace{1em} Ensemble[m] = $W$;
13: end for

**DISCUSSION**

The exercises carried out so far have highlighted a number of (stylized) facts concerning the performance of the three algorithms tested: in what follows, we will
briefly sum them up.

Newton’s method. Overall, Newton’s method is very accurate - often, the most accurate one - in reproducing both the binary and the weighted constraints; moreover, it represents the only viable alternative when the most complicated models are considered (i.e. the UECM and the DECM, respectively defined by a system of $2N$ and $4N$ coupled, non-linear equations). However, the time required to run Newton’s method on a given model seems to be quite dependent on the network size, especially whenever the corresponding system of equations cannot be reduced - see the case of the undirected CReM, run on the Bitcoin Lightning Network. Since one of the reasons affecting the bad scaling of Newton’s method with the network size is the evaluation of the Hessian matrix defining a given model, this algorithm has to be preferred for largely reducible networks.

Quasi-Newton method. For all the networks considered here, the quasi-Newton method we have implemented is nothing else than the diagonal version of the traditional Newton’s method. Even if this choice greatly reduces the number of entries of the Hessian matrix which are needed (i.e. just $N$ elements for the undirected version of the CReM, $2N$ elements for the UECM and the directed version of the CReM and $4N$ elements for the DECM) dimensionality may still represent an issue to achieve fast convergence. Moreover, since the diagonal approximation of the Hessian matrix is not necessarily always a good one, the quasi-Newton method may require more time than Newton’s one to achieve the same level of accuracy in reproducing the constraints. However, when such an approximation is a good one, the ‘regime’ in which the quasi-Newton method outperforms the competitors seems to be the one of small, non-reducible networks (e.g. see the results concerning the DBCM run on the WTW) - although, in cases like these, Newton’s method may still be a strong competitor.

Fixed-point method. From a purely theoretical point of view, the fixed-point recipe is the fastest one, since the time required to evaluate the generic $n$-th step is (only) due to the evaluation of the model-specific map at the $(n−1)$-th iteration. Strictly speaking, however, this holds true for a single step: if the number of steps required for convergence is large, in fact, the total amount of time required by the fixed-point method can be large as well. Overall, however, this algorithm has to be preferred for large, non-reducible networks: this is the case of the (undirected version of the) CReM, run on the 8-th snapshot of the Bitcoin Lightning Network (i.e. day 17-07-19) and requiring a little more than one minute to achieve an accuracy of MRDE$_{\text{fixed-point}} \gtrsim 10^{-10}$ and MRSE$_{\text{fixed-point}} \simeq 10^{-1}$; naturally, the method is not as accurate as Newton’s one, for which MRDE$_{\text{Newton}} \gtrsim 10^{-12}$ and MRSE$_{\text{Newton}} \simeq 10^{-6}$ but is much faster as Newton’s algorithm requires $\approx 1188$ seconds to converge.

The ‘NEMTROPY’ Python package. As an additional result, we release a comprehensive package, coded in Python, that implements the three aforementioned algorithms on all the ERGMs considered in the present work. Its name is ‘NEMTROPY’, an acronym standing for ‘Network Entropy Maximization: a Toolbox Running On Python’, and is freely downloadable at the following URL: https://pypi.org/project/NEMtropy/.

Alternative techniques to improve accuracy and speed have been tested as well, as the one of coupling two of the algorithms considered above. In particular, we have tried to solve the (undirected version of the) CReM by running the fixed-point algorithm and using the solution of the latter as input for the quasi-Newton method. The results are reported in Table VIII as they clearly show, the coupled algorithm is indeed more accurate that the single methods composing it and much faster than the quasi-Newton one (for some snapshots, more accurate and even faster than Newton’s method). Techniques like these, are, in general, useful to individuate better initial conditions than the completely random ones: a first run of the fastest method may be, in fact, useful to direct the most accurate algorithm towards the (best) solution. This is indeed the case, upon considering that the quasi-Newton method, now, stops because the condition $\|\nabla \mathcal{L}(\bar{\theta})\|_2 \leq 10^{-8}$ is satisfied - and not for having reached the limit of 10000 steps.

We would like to end the discussion about the results presented in this contribution by explicitly mentioning a circumstance that is frequently met when studying economic and financial networks. When considering systems like these, the information about the number of neighbours of each node is typically not accessible: as a consequence, the models constraining both binary and weighted information cannot be employed as they have presented in this contribution.

Alternatives exist and rest upon the existence of some kind of relationship between binary and weighted constraints. In the case of undirected networks, such a relationship is usually written as

$$e^{-\theta_i} = \sqrt{2} s_i$$

and establishes that the Lagrange multipliers controlling for the degrees are linearly proportional to the strengths. If this is the case (or a valid reason exists for this to be the case), the expression for the probability that any two nodes are connected becomes
Fixed-point + Quasi-Newton

|                | MRDE   | MADE   | MRSE   | Time (s) |
|----------------|--------|--------|--------|----------|
| BLN 24-01-18   | \(\approx 2 \cdot 10^{-7}\) | \(\approx 1.1 \cdot 10^{-7}\) | \(\approx 1 \cdot 10^{-9}\) | \(\approx 0.1\) |
| BLN 25-02-18   | \(\approx 1.3 \cdot 10^{-7}\) | \(\approx 1.5 \cdot 10^{-9}\) | \(\approx 1 \cdot 10^{-9}\) | \(\approx 1.6\) |
| BLN 30-03-18   | \(\approx 1 \cdot 10^{-7}\) | \(\approx 1.3 \cdot 10^{-10}\) | \(\approx 7.8 \cdot 10^{-9}\) | \(\approx 2.2\) |
| BLN 13-07-18   | \(\approx 7.5 \cdot 10^{-8}\) | \(\approx 4.2 \cdot 10^{-8}\) | \(\approx 5.3 \cdot 10^{-8}\) | \(\approx 200\) |
| BLN 19-12-18   | \(\approx 7.5 \cdot 10^{-8}\) | \(\approx 1.7 \cdot 10^{-8}\) | \(\approx 1.1 \cdot 10^{-9}\) | \(\approx 7\) |
| BLN 30-01-19   | \(\approx 6.2 \cdot 10^{-8}\) | \(\approx 1.8 \cdot 10^{-8}\) | \(\approx 4.1 \cdot 10^{-9}\) | \(\approx 614\) |
| BLN 01-03-19   | \(\approx 5.7 \cdot 10^{-8}\) | \(\approx 5.4 \cdot 10^{-8}\) | \(\approx 9.1 \cdot 10^{-9}\) | \(\approx 961\) |
| BLN 17-07-19   | \(\approx 4.9 \cdot 10^{-8}\) | \(\approx 1.3 \cdot 10^{-8}\) | \(\approx 3.5 \cdot 10^{-9}\) | \(\approx 3350\) |

TABLE VIII. Performance of the algorithm coupling fixed-point and quasi-Newton to solve the system of equations defining the (undirected version of the) CReM, on a set of real-world WUNs. The algorithm stops because the condition \(|\nabla \mathcal{L}(\bar{\theta})|_2 \leq 10^{-8}\) is satisfied. As the results reveal, it is more accurate that the single methods composing it and much faster than the quasi-Newton one - for some snapshots, more accurate and even faster than Newton’s method. Only the results corresponding to the best choice of initial conditions are reported.

\[
p_{ij}^{dcGM} = \frac{z s_i s_j}{1 + z s_i s_j} \quad \forall \: i < j
\]

the acronym standing for degree-corrected Gravity Model [44]. The (only) unknown parameter \(z\) must be numerically estimated by employing some kind of topological information; this is usually represented by (a proxy of) the network link density, used to instantiate the (only) likelihood condition

\[
L(A^*) = \langle L \rangle = \sum_{i=1}^{N} \sum_{\substack{j=1 \atop (j<i)}}^{N} \frac{z s_i s_j}{1 + z s_i s_j}; \quad (81)
\]

once the equation above has been solved, the set of coefficients \(\{p_{ij}^{dcGM}\}_{i,j=1}^{N}\) can be either employed 1) to, first, estimate the degrees and, then, solve the UECM [45] or 2) within the CReM framework, via the identification \(f_{ij} \equiv p_{ij}^{dcGM}\), to estimate the parameters controlling for the weighted constraints.

CONCLUSIONS

The definition and correct implementation of null models is a crucial issue in network analysis: the present contribution focuses on (a subset of) the ones constituting the so-called ERG framework - a choice driven by the evidence that they are the most commonly employed ones for tasks as different as network reconstruction, pattern detection, graph enumeration. The optimization of the likelihood function associated to them is, however, still problematic since it involves the resolution of large systems of coupled, non-linear equations.

Here, we have implemented and compared three algorithms for numerical optimization, with the aim of finding the one performing best (i.e. being both accurate and fast) for each model. What emerges from our results is that there is no a unique method which is both accurate and fast for all models on all configurations: under this respect, performance is just a trade-off between accuracy and speed. However, some general conclusions can still be drawn.

Newton’s method is the one requiring the largest amount of information per step (in fact, the entire Hessian matrix has to be evaluated at each iteration): hence, it is the most accurate one but, for the same reason, often the one characterized by the slowest performance. A major drawback of Newton’s method is that of scaling very badly with the network size.

At the opposite extreme lies the fixed-point algorithm, theoretically the fastest one but, often, among the least accurate ones (at least, for what concerns the weighted constraints); the performance if the quasi-Newton method often lies in-between the performances of the two methods above, by achieving an accuracy that is larger than the one achieved by the fixed-point algorithm, requiring less time that the one needed by Newton’s method.

Overall, while Newton’s method seems to perform best on either relatively small or greatly reducible networks, the fixed-point method must be preferred for large, non-reducible configurations. Deviations from this (over simplified) picture are, however, clearly visible.

Future work concerns the application of the aforementioned three numerical recipes to the models that have not found place here. For what concerns the set of purely binary constraints, the ones defining the Reciprocal Binary Configuration Model (RBCM) [15] deserve to be mentioned.

For what concerns the ‘mixed’ constraints, instead, the CReM framework is versatile enough to accommodate a quite large number of variants. In the present work, we have ‘limited’ ourselves to combine the UBCM and the DBCM with (conditional) distributions of continuous weights: a first, obvious, generalization is that of
considering the discrete versions of such models, defined by the positions
\[
\langle w_{ij}\rangle_{d\text{-CREM}} = \frac{f_{ij}}{1 - e^{-\beta_j - \beta_i}}
\] (82)
with \(f_{ij} = p_{ij}^{UBCM}\) and
\[
\langle w_{ij}\rangle_{d\text{-CREM}} = \frac{f_{ij}}{1 - e^{-\gamma_j - \delta_i}}
\] (83)
with \(f_{ij} = p_{ij}^{DBCM}\); a second one concerns the continuous versions of the UECM and of the DECM, respectively defined by the positions \[19\]
\[
p_{ij}^{UECM} = \frac{e^{-\alpha_i - \alpha_j}}{\beta_i + \beta_j + e^{-\alpha_i - \alpha_j}},
\] (84)
and
\[
p_{ij}^{DECM} = \frac{e^{-\alpha_i - \beta_j}}{\gamma_i + \delta_j + e^{-\alpha_i - \beta_j}}.
\] (85)

AUTHORS CONTRIBUTIONS

FS, GC, MZ, TS developed the methods and designed the research. NV, MB, EM, GT performed the analysis (NV: DBCM, DECM, RBCM; MB: BiCM; EM: UBCM, UECM, CREM; GT: preliminary version of the BiCM). FS, GC, MZ, TS wrote the manuscript. All authors reviewed and approved the manuscript.

ACKNOWLEDGEMENTS

FS and TS acknowledge support from the European project SoBigData++ (GA. 871042). GC and MZ acknowledge support from the PAI project ORCS (‘Optimized Reconstruction of Complex networkS’), funded by the IMT School for Advanced Studies Lucca. FS also acknowledges support from the Italian ‘Programma di Attività Integrata’ (PAI) project ‘TOol for Fighting FakEs’ (TOFFE), funded by IMT School for Advanced Studies Lucca.

APPENDIX A: COMPUTING THE HESSIAN MATRIX

As we showed in the main text, the Hessian matrix of our likelihood function is ‘minus’ the covariance matrix of the constraints, i.e.
\[
H_{ij} = \frac{\partial^2 \mathcal{L}(\tilde{\theta})}{\partial \theta_i \partial \theta_j} = -\mathrm{Cov}[C_i, C_j], \quad i, j = 1 \ldots M;
\] (86)
interestingly, a variety of alternative methods exists to explicitly calculate the generic entry \(H_{ij}\), i.e. 1) taking the second derivatives of the likelihood function characterizing the method under analysis, 2) taking the first derivatives of the expectation values of the constraints characterizing the method under analysis, 3) calculating the moments of the pair-specific probability distributions characterizing each method.

UBCM: binary undirected graphs with given degree sequence

The Hessian matrix for the UBCM is an \(N \times N\) symmetric table with entries reading
\[
\begin{align*}
H_{UBCM} &= \left\{ \begin{array}{ll}
\text{Var}[k_i] = \sum_{j=1}^{N} p_{ij} (1 - p_{ij}), \quad \forall i \\
\text{Cov}[k_i, k_j] = p_{ij} (1 - p_{ij}), \quad \forall i \neq j
\end{array} \right. \\
\end{align*}
\] (87)
where \(p_{ij} = p_{ij}^{UBCM}\). Notice that \(\text{Var}[k_i] = \sum_{j=1}^{N} \text{Cov}[k_i, k_j], \forall i\).

DBCM: binary directed graphs with given in-degree and out-degree sequences

The Hessian matrix for the DBCM is a \(2N \times 2N\) symmetric table that can be further subdivided into four \(N \times N\) blocks whose entries read
\[
\begin{align*}
H_{DBCM} &= \left\{ \begin{array}{ll}
\text{Var}[k_i^{\text{out}}] = \sum_{j=1}^{N} p_{ij} (1 - p_{ij}), \quad \forall i \\
\text{Var}[k_i^{\text{in}}] = \sum_{j=1}^{N} p_{ji} (1 - p_{ji}), \quad \forall i \\
\text{Cov}[k_i^{\text{out}}, k_j^{\text{in}}] = p_{ij} (1 - p_{ij}), \quad \forall i \neq j \\
\text{Cov}[k_j^{\text{out}}, k_i^{\text{in}}] = p_{ji} (1 - p_{ji}), \quad \forall i \neq j
\end{array} \right. \\
\end{align*}
\] (88)
while \(\text{Cov}[k_i^{\text{out}}, k_i^{\text{in}}] = \text{Cov}[k_i^{\text{in}}, k_i^{\text{out}}] = 0\) and \(p_{ij} = p_{ij}^{DBCM}\). Notice that the Hessian matrix of the BiCM mimicks the DBCM one, the only difference being that the probability coefficients are now indexed by \(i\) and \(\alpha\): for example, in the BiCM case, one has that \(\text{Cov}[k_i, d_\alpha] = p_{\alpha i} (1 - p_{\alpha i}), \forall i, \alpha\).
UECM: weighted undirected graphs with given strengths and degrees

The Hessian matrix for the UECM is a 2N × 2N symmetric table that can be further subdivided into four blocks (each of which with dimensions N × N). In order to save space, the expressions indexed by the single subscript i will be assumed as being valid ∀i, while the ones indexed by a double subscript i, j will be assumed as being valid ∀i ≠ j. The entries of the diagonal blocks read

\[
H_{\text{UECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial \alpha_i^2} = \text{Var}[k_i] = \sum_{j=1}^{N} p_{ij} (1 - p_{ij}) \\
\frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j} = \text{Cov}[k_i, k_j] = p_{ij} (1 - p_{ij}) 
\end{cases}
\]

and

\[
H_{\text{UECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial \beta_i^2} = \text{Var}[s_i] = \sum_{j=1}^{N} p_{ij} (1 - p_{ij}) \\
\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} = \text{Cov}[s_i, s_j] = p_{ij} (1 - p_{ij}) 
\end{cases}
\]

with \( p_{ij} \equiv p_{ij}^{\text{UECM}} \).

DECDec: weighted directed graphs with given strengths and degrees

The Hessian matrix for the DECM is a 4N × 4N symmetric table that can be further subdivided into four blocks (each of which with dimensions N × N). As for the UECM, in order to save space, the expressions indexed by the single subscript i will be assumed as being valid ∀i, while the ones indexed by a double subscript i, j will be assumed as being valid ∀i ≠ j. The entries of the diagonal blocks read

\[
H_{\text{DECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial \alpha_i^2} = \text{Var}[s_i^{\text{out}}] = \sum_{j=1}^{N} p_{ij} (1 - p_{ij}) \\
\frac{\partial^2 L}{\partial \alpha_i \partial \alpha_j} = \text{Cov}[s_i^{\text{out}}, s_j^{\text{out}}] = 0 
\end{cases}
\]

and

\[
H_{\text{DECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial \beta_i^2} = \text{Var}[s_i^{\text{in}}] = \sum_{j=1}^{N} p_{ij} (1 - p_{ij}) \\
\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} = \text{Cov}[s_i^{\text{in}}, s_j^{\text{in}}] = 0 
\end{cases}
\]

and

\[
H_{\text{DECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial r_i \partial r_j} = \text{Cov}[r_i^{\text{out}}, r_j^{\text{out}}] = 0 \\
\frac{\partial^2 L}{\partial r_i \partial s_j} = \text{Cov}[r_i^{\text{out}}, s_j^{\text{out}}] = 0 
\end{cases}
\]

and

\[
H_{\text{DECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial r_i \partial r_j} = \text{Cov}[r_i^{\text{in}}, s_j^{\text{in}}] = 0 \\
\frac{\partial^2 L}{\partial r_i \partial s_j} = \text{Cov}[r_i^{\text{in}}, s_j^{\text{in}}] = 0 
\end{cases}
\]

and

\[
H_{\text{DECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial \gamma_i \partial \gamma_j} = \text{Cov}[\gamma_i^{\text{out}}, \gamma_j^{\text{out}}] = 0 \\
\frac{\partial^2 L}{\partial \gamma_i \partial \delta_j} = \text{Cov}[\gamma_i^{\text{out}}, \delta_j^{\text{out}}] = 0 
\end{cases}
\]

and

\[
H_{\text{DECM}} = \begin{cases} 
\frac{\partial^2 L}{\partial \gamma_i \partial \gamma_j} = \text{Cov}[\gamma_i^{\text{in}}, \gamma_j^{\text{in}}] = 0 \\
\frac{\partial^2 L}{\partial \gamma_i \partial \delta_j} = \text{Cov}[\gamma_i^{\text{in}}, \delta_j^{\text{in}}] = 0 
\end{cases}
\]

with \( p_{ij} \equiv p_{ij}^{\text{DECM}} \).
Two-step models for undirected and directed networks

The Hessian matrix for the undirected two-step model considered here is an \( N \times N \) symmetric table reading

\[
H_{\text{und}}^\text{CReM} = \begin{cases} 
\text{Var}[s_i] = \sum_{j=1}^{N-1} \frac{f_{ij}}{(\alpha_i + \beta_j)^2}, & \forall i \\
\text{Cov}[s_i, s_j] = \frac{f_{ij}}{(\alpha_i + \beta_j)^2}, & \forall i \neq j \end{cases}
\]

where \( f_{ij} \) is given. In the directed case, instead, the Hessian matrix for the two-step model considered here is a \( 2N \times 2N \) symmetric table that can be further subdivided into four \( N \times N \) blocks whose entries read

\[
H_{\text{dir}}^\text{CReM} = \begin{cases} 
\text{Var}[s_i^\text{out}] = \sum_{j=1}^{N-1} \frac{f_{ij}}{(\alpha_i + \beta_j)^2}, & \forall i \\
\text{Var}[s_i^\text{in}] = \sum_{j=1}^{N-1} \frac{f_{ij}}{(\alpha_i + \beta_j)^2}, & \forall i \\
\text{Cov}[s_i^\text{out}, s_j^\text{in}] = \frac{f_{ij}}{(\alpha_i + \beta_j)^2}, & \forall i \neq j 
\end{cases}
\]

while \( \text{Cov}[s_i^\text{out}, s_i^\text{out}] = \text{Cov}[s_i^\text{in}, s_i^\text{in}] = 0 \) and \( f_{ij} \) is given.

APPENDIX B: A NOTE ON THE CHANGE OF VARIABLES

In all methods we will considered in the present work, the variable \( \theta_i \) appears in the optimality conditions only through negative exponential functions: it is therefore tempting to perform the change of variable \( x_i = e^{-\theta_i} \). Although this is often performed in the literature, one cannot guarantee that the new optimization problem remains convex: in fact, simple examples can be provided for which convexity is lost. This has several consequences, e.g. 1) convergence to the global maximum is no longer guaranteed (since the existence of a global maximum is no longer guaranteed as well), 2) extra-care is needed to guarantee that the Hessian matrix \( H \) employed in our algorithms is negative definite. While problem 2) introduces additional complexity only for Newton’s method, problem 1) is more serious from a theoretical point of view.

Let us now address problem 1) in more detail. First, it is possible to prove that any stationary point for \( \mathcal{L}(\bar{x}) \) satisfies the optimality conditions for \( \mathcal{L}(\bar{\theta}) \) as well. In fact, the application of the ‘chain rule’ leads to recover the set of relationships

\[
\frac{\partial \mathcal{L}(\bar{\theta})}{\partial \theta_i} = \frac{\partial x_i \partial \mathcal{L}(\bar{x})}{\partial \theta_i \partial x_i} = -x_i \frac{\partial \mathcal{L}(\bar{x})}{\partial x_i}, \quad i = 1 \ldots M;
\]

notice that requiring \( \nabla_\theta \mathcal{L}(\bar{\theta}) = 0 \) leads to require that either \( \nabla_{x_i} \mathcal{L}(\bar{x}) = 0 \) or \( x_i = 0 \). As the second eventual-
ity precisely identifies isolated nodes (i.e. the nodes for which the constraint \( C_i(G^*) \), controlled by the multiplier \( \theta_i \), is 0), one can get rid of it by explicitly removing the corresponding addenda from the likelihood function.

For what concerns convexity, let us explicitly calculate the Hessian matrix for the set of variables \( \{x_i\}_{i=1}^M \). In formulas,

\[
\frac{\partial^2 \mathcal{L}(\bar{x})}{\partial x_i^2} = e^{2\theta_i} \left( \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta_i^2} + \frac{\partial \mathcal{L}(\bar{\theta})}{\partial \theta_i} \right), \quad i = 1 \ldots M,
\]

\[
\frac{\partial^2 \mathcal{L}(\bar{x})}{\partial x_i \partial x_j} = e^{\theta_i + \theta_j} \left( \frac{\partial^2 \mathcal{L}(\bar{\theta})}{\partial \theta_i \partial \theta_j} \right), \quad \forall i \neq j
\]

according to the ‘chain rule’ for second-order derivatives. More compactly,

\[
H_{\mathcal{L}(\bar{x})} = e^\theta \circ \left( -\text{Cov}[C_i, C_j] + \text{I} \cdot \nabla_\theta \mathcal{L}(\bar{\theta}) \right)
\]

where \( \text{I} \) is the identity matrix, the generic entry of the matrix \( e^\theta \) reads \( [e^\theta]_{ij} = e^{\theta_i + \theta_j}, \forall i, j \) and the symbol ‘\( \circ \)’ indicates the Hadamard (i.e. element-wise) product of matrices. In general, the expression above defines an indefinite matrix, i.e. a neither positive nor negative (semi)definite one.

APPENDIX C: FIXED POINT METHOD IN THE MULTIVARIATE CASE

Equation (21) can be written as

\[
\theta_i^{(n)} = G_i(\bar{\theta}^{(n-1)}), \quad i = 1 \ldots N;
\]

for the sake of illustration, let us discuss it for the UBCM case. In this particular case, the set of equations above can be rewritten as

\[
\theta_i^{(n)} = -\ln \left[ \frac{k_i(A^*)}{\sum_{j=1}^{N} \frac{e^{-\theta_j^{(n-1)}}}{1 + e^{-\theta_j^{(n-1)}}}} \right], \quad i = 1 \ldots N.
\]

Since all components of the map \( G \) are continuous on \( \mathbb{R}^N \), the map itself is continuous on \( \mathbb{R}^N \). Hence, a fixed point exists. Let us now consider its Jacobian matrix and check the magnitude of its elements. In the UBCM case, one finds that
following relationships hold

\[ \frac{\partial G_i}{\partial \theta_i} = \sum_{j \neq i}^{N} \frac{e^{-\theta_j}}{1 + e^{-\theta_i - \theta_j}} = \sum_{j \neq i}^{N} \left( \frac{e^{-\theta_j}}{1 + e^{-\theta_i - \theta_j}} \right)^2 = 1 - \text{Var}[k_i] = 1 - \sum_{j \neq i} \text{Cov}[k_i, k_j], \quad \forall i \]

(109)

and

\[ \frac{\partial G_i}{\partial \theta_j} = -\frac{e^{-\theta_j}}{(1+e^{-\theta_i-\theta_j})^2} = -\frac{e^{-\theta_i-\theta_j}}{(1+e^{-\theta_i-\theta_j})^2} = -\text{Cov}[k_i, k_j], \quad \forall i, j. \]

(110)

Let us notice that 1) each element of the Jacobian matrix is a continuous function \( \mathbb{R}^N \rightarrow \mathbb{R} \) and that 2) the following relationships hold

\[ \left| \frac{\partial G_i}{\partial \theta_j} \right| \leq 1, \quad \forall i, j; \]

(111)

unfortunately, however, when multivariate functions are considered, the set of conditions above is not enough to ensure convergence to the fixed point for any choice of the initial value of the parameters. What is needed to be checked is the condition \( \| J_G(\hat{\theta}) \| < 1 \), with \( J \) indicating the Jacobian of the map (i.e. the matrix of the first, partial derivatives above) and \( \| \cdot \| \) any natural matrix norm: the validity of such a condition has been numerically verified case by case.

\[ ^{\ast} \text{tiziano.squartini@imtlucca.it} \]

[1] M.E.J. Newman, *Networks: an introduction*, Oxford University Press (2010).

[2] V. Colizza, A. Barrat, M. Barthelemy, A. Vespignani, *The role of the airline transportation network in the prediction and predictability of global epidemics*, Proceedings of the National Academy of Sciences 103 (7), 2015-2020 (2006).

[3] A. Barrat, M. Barthelemy, A. Vespignani, *Dynamical processes on complex networks*, Cambridge University Press (2008).

[4] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, A. Vespignani, *Epidemic processes in complex networks*, Rev. Mod. Phys. 87, 925 (2015).

[5] C. Castellano, S. Fortunato, V. Loreto, *Statistical physics of social dynamics*, Rev. Mod. Phys. 81, 591 (2009).

[6] T. Squartini, I. van Lelyveld, D. Garlaschelli, *Early-warning signals of topological collapse in interbank networks*, Sci. Rep. 3, 3357 (2013).

[7] G. Cimini, T. Squartini, F. Saracco, D. Garlaschelli, A. Gabrielli, G. Caldarelli, *The statistical physics of real-world networks*, Nat. Rev. Phys. 1 (1), 58-71 (2019).

[8] S. Maslov and K. Sneppen, *Specificity and stability in topology of protein networks*, Science 296 (5569), 910-913 (2002).

[9] A.C.C. Coolen, A. De Martino, A. Annibale, *Constrained Markovian dynamics of random graphs*, J. Stat. Phys. 136, 1035-1067 (2009).

[10] E.S. Roberts, A.C.C. Coolen, *Unbiased degree-preserving randomization of directed binary networks*, Phys. Rev. E 85 (4), 046103 (2012).

[11] Y. Artzy-Randrup, L. Stone, *Generating uniformly distributed random networks*, Phys. Rev. E 72 (5), 056708 (2005).

[12] C.I. Del Genio, H. Kim, Z. Toroczkai, K.E. Bassler, C.I. Del Genio, H. Kim, Z. Toroczkai, K.E. Bassler, PLoS One 5 (4), e10012 (2010).

[13] H. Kim, C.I. Del Genio, K.E. Bassler, Z. Toroczkai, *Constructing and sampling directed graphs with given degree sequences*, New J. Phys. 14, 023012 (2012).

[14] J. Blitzstein, P. Diaconis, *A sequential importance sampling algorithm for generating random graphs with prescribed degrees*, Internet Mathematics 6 (4), 489-522 (2011).

[15] T. Squartini, D. Garlaschelli, *Analytical maximum-likelihood method to detect patterns in real networks*, New J. Phys. 13. 083001 (2011).

[16] J. Park, M.E.J. Newman, *Statistical mechanics of networks*, Phys. Rev. E 70 (6), 066117 (2004).

[17] G. Bianconi, *The entropy of randomized network ensembles*, Europhys. Lett. 81 (2), 28005 (2007).

[18] A. Fronczak, P. Fronczak, J.A. Holyst, *Fluctuation-dissipation relations in complex networks*, Phys. Rev. E 73 (1), 016108 (2006).

[19] A. Gabrielli, R. Mastrandrea, G. Caldarelli, G. Cimini, *Grand canonical ensemble of weighted networks*, Phys. Rev. E 99 (3), 030301(R) (2019).

[20] A. Fronczak, *Exponential random graph models in Encyclopedia of Social Network Analysis and Mining*, Springer (edited by R. Alhajj and J. Rokne) (2014).

[21] E.T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev. 106 (4), 620-630 (1957).

[22] N. Dianati, *A maximum entropy approach to separating noise from signal in bimodal affiliation networks*, arXiv:2005.00114 (2020).

[23] N. Vallarano, C. Tessone, T. Squartini, *Bitcoin Transaction Networks: an overview of recent results*, arXiv:2005.00114 (2020).

[24] D. Garlaschelli, M.I. Loffredo, *Maximum likelihood: Extracting unbiased information from complex networks*, Phys. Rev. E 78 (1), 015101(R) (2008).

[25] J. Nocedal, S.J. Wright, *Numerical Optimization*, Springer (2006).

[26] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press (2004).

[27] F. Chung, L. Lu, *Connected components in random graphs with given expected degree sequences*, Ann. Combinatorics, 6. 125-145 (2002).

[28] K. Oshio, Y. Iwasaki, S. Morita, Y. Osana, S. Gomi, E. Akiyama, K. Omata, K. Oka and K. Kawamura, *Database of Sentic Connectivity of C. elegans*, Tech. Rep. of CCEP, Keio Future 3, (Keio University, 2003).

[29] V. Colizza, R. Pastor-Satorras and A. Vespignani, *Re-
action–diffusion processes and metapopulation models in heterogeneous networks, Nat. Phys. 3, 276-282 (2007).

[30] http://dip.doe-mbi.ucla.edu/dip/Main.cgi

[31] V. Colizza, A. Flammini, M.A. Serrano and A. Vespignani, Detecting rich-club ordering in complex networks, Nat. Phys. 2, 110-115 (2006).

[32] J.-H. Lin, K. Primicerio, T. Squartini, C. Decker, C.J.Tessone, Lightning Network: a second path towards centralisation of the Bitcoin economy, New J. Phys. 22, 083022 (2020).

[33] J.C. Miller, A. Hagberg, Efficient generation of networks with given expected degrees, LNCS 6732, 115-126, Springer (edited by A. Frieze, P. Horn and P. Pralat) (2011).

[34] T. Squartini, G. Fagiolo, D. Garlaschelli, Randomizing world trade. I. A binary network analysis, Phys. Rev. E 84, 046117 (2011).

[35] A. Bovet, C. Campajola, F. Mottes, V. Restocchi, N. Val larano, T. Squartini, C.J. Tessone, The evolving biaisons between the transaction networks of Bitcoin and its price dynamics, arXiv:1907.03577 (2019).

[36] G. Caldarelli, R. de Nicola, M. Petrocchi, M. Pratelli, F. Saracco, Flow of online misinformation during the peak of the COVID-19 pandemic in Italy, arXiv:2010.01913 (2020).

[37] F. Saracco, R. Di Clemente, A. Gabrielli, T. Squartini, Randomizing bipartite networks: the case of the World Trade Web, Sci. Rep. 5, 10595 (2015).

[38] R. Mastrandrea, T. Squartini, G. Fagiolo, D. Garlaschelli, Intensive and extensive biases in economic networks: reconstructing world trade, New J. Phys. 16, 043022 (2014).

[39] D. Garlaschelli, M.I. Loffredo, Generalized bose-fermi statistics and structural correlations in weighted networks, Phys. Rev. Lett. 102 (3), 038701 (2009).

[40] R. Mastrandrea, T. Squartini, G. Fagiolo, D. Garlaschelli, Reconstructing the world trade multiplex: the role of intensive and extensive biases, Phys. Rev. E 90 (6), 062804 (2014).

[41] K. Gleditsch, Expanded trade and GDP data, J. Conflict Resol. 46, 712-24 (2002).

[42] G. Iori, G. De Masi, O. Vasile Precup, G. Gabbi, G. Caldarelli, A network analysis of the Italian overnight money market, J. Econ. Dyn. Control 32 (1), 259-278 (2006).

[43] F. Parisi, T. Squartini, D. Garlaschelli, A faster horse on a safer trail: generalized inference for the efficient reconstruction of weighted networks, New J. Phys. 22, 053053 (2020).

[44] G. Cimini, T. Squartini, A. Gabrielli, D. Garlaschelli, Estimating topological properties of weighted networks from limited information, Phys. Rev. E 92, 040802 (2015).

[45] G. Cimini, T. Squartini, A. Gabrielli, D. Garlaschelli, Systemic risk analysis on reconstructed economic and financial networks, Sci. Rep. 5, 15758 (2015).