

Quantifying the Multi-Scale Performance of Network Inference Algorithms¹

Chris. J. Oates² and Simon E. F. Spencer
Department of Statistics, University of Warwick, Coventry, UK

Richard Amos
The MathWorks, Cambridge, UK.

March 3, 2014

¹Running title: Multi-Scale Scores for Network Inference

²Corresponding author: c.oates@warwick.ac.uk

Abstract

Graphical models are widely used to study complex multivariate biological systems. Network inference algorithms aim to reverse-engineer such models from noisy experimental data. It is common to assess such algorithms using techniques from classifier analysis. These metrics, based on ability to correctly infer individual edges, possess a number of appealing features including invariance to rank-preserving transformation. However, regulation in biological systems occurs on multiple scales and existing metrics do not take into account the correctness of higher-order network structure.

In this paper novel performance scores are presented that share the appealing properties of existing scores, whilst capturing ability to uncover regulation on multiple scales. Theoretical results confirm that performance of a network inference algorithm depends crucially on the scale at which inferences are to be made; in particular strong local performance does not guarantee accurate reconstruction of higher-order topology. Applying these scores to a large corpus of data from the DREAM5 challenge, we undertake a data-driven assessment of estimator performance. We find that the “wisdom of crowds” network, that demonstrated superior local performance in the DREAM5 challenge, is also among the best performing methodologies for inference of regulation on multiple length scales.

MATLAB R2013b code `net_assess` is provided as Supplement.

Key words: Performance assessment, multi-scale scores, network inference.

1 Introduction

Graphical representations of complex multivariate systems are increasingly prevalent within systems biology. In general a graph or *network* $G = (V, E)$ is characterised by a set V of vertices (typically associated with molecular species) and a set $E \subseteq V \times V$ of edges, whose interpretation will be context-specific. In many situations the edge set or *topology* $E \equiv E(G)$ is taken to imply conditional independence relationships between species in V (Pearl, 2000). For fixed and known vertex set V , the data-driven characterisation of network topology is commonly referred to as *network inference*.

In the last decade many approaches to network inference have been proposed and exploited for several different purposes (Oates and Mukherjee, 2012a). In some settings it is desired to infer single edges with high precision (e.g. Hill, 2012), whereas in other applications it is desired to infer global connectivity, such as subnetworks and clusters (e.g. Breitkreutz *et al.*, 2012). In cellular signalling systems, the scientific goal is often to identify a set of upstream regulators for a given target, each of which is a candidate for therapeutic intervention designed to modulate activity of the target (Morrison *et al.*, 2005; Winter *et al.*, 2012). The output of network inference algorithms are increasingly used to inform the design of experiments (Nelander *et al.*, 2008; Hill, 2012) and may soon enter into the design of clinical trials (Chuang *et al.*, 2007; Heiser *et al.*, 2012). It is therefore important to establish which network inference algorithms work best for each of these distinct scientific goals.

Assessment of algorithm performance can be achieved *in silico* by comparing inferred networks to known data-generating networks. It can also be achieved using data obtained *in vitro*; however this requires that the underlying biology is either known by design (Cantone *et al.*, 2009), well-characterised by interventional experiments (Maathuis *et al.*, 2010), or estimated from larger corpora of data (Weile *et al.*, 2012). In either case an estimated network \hat{G} , typically represented as a weighted adjacency matrix, is compared against a known or assumed benchmark network G . Community-wide blind testing of network inference algorithms is performed at the regular DREAM challenges (see <http://www.the-dream-project.org/>; Marbach *et al.*, 2012; Prill *et al.*, 2010).

There is broad agreement in the network inference literature regarding the selection of suitable performance scores (described below), facilitating the comparison of often disparate methodologies across publications. In this literature, the quality of an estimated network \hat{G} with respect to a benchmark G is assessed using techniques from classifier analysis. That is, each possible edge $(i, j) \in V \times V$ has an associated class label $Z(i, j) = \mathbb{I}\{(i, j) \in E(G)\}$, where \mathbb{I} is the indicator function. A network estimator \hat{G} may then be seen as an attempt to estimate $Z(i, j)$ for each pair (i, j) . Two of the main performance scores from classifier analysis are area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR), though alternative performance scores for classification also exist (e.g. Drummond and Holte (2004)). These scores, formally defined in Sec. 2.4, are based on *confusion matrices* of true/false positive/negative counts and represent essentially the only way to quantify performance at a local level (i.e. based on individual edges). At present, performance assessment in the network inference literature does not typically distinguish between the various scientific purposes for which network inference algorithms

are to be used. Yet network inference algorithms are now frequently employed to perform diverse tasks, including identifying single edges with high precision (Hill, 2012), eliciting network motifs such as cliques (Wang *et al.*, 2008; Feizi *et al.*, 2013) or learning a coherent global topology such as connected components (Breitkreutz *et al.*, 2012).

Whilst performance for local (i.e. edge-by-edge) recovery is now standardised, there has been comparatively little attention afforded to performance scores that capture ability to recover higher-order features such as cliques, motifs and connectivity. Recent studies, including Banerjee and Jost (2009); Jurman *et al.* (2011a,b), proposed to employ spectral distances as a basis for comparing between two networks on multiple length scales. In this article we present several additional multi-scale scores (MSSs) for network inference algorithms, each of which reflects ability to solve a particular class of inference problem. Much of the popularity of existing scores derives from their objectivity, interpretability and invariance to rank-preserving transformation. Unlike previous studies, we restrict attention only to MSS that satisfy these desiderata.

The remainder of this paper proceeds as follows: In Section 2 we formulate the assessment problem, state our desiderata and present novel performance scores that satisfy these requirements whilst capturing aspects of network reconstruction on multiple scales. Using a large corpus of estimated and benchmark networks from the DREAM5 Challenge in Section 3, we survey estimator performance and conduct an objective, data-driven examination of the statistical power of each MSS. The proposed MSSs provide evidence that the “wisdom of crowds” approach, that demonstrated superior (local) performance in the DREAM5 challenge, also offers gains on multiple length scales. Sections 4 and 5 provide a discussion of our proposals and suggest directions for future work. MATLAB R2013b code `net_assess` is provided in the Supplement, to accelerate the dissemination of ideas discussed herein.

2 Methods

We proceed as follows: Sections 2.1 and 2.2 clarify the context of the assessment problem for network inference algorithms and list certain desiderata that have contributed to the popularity of local scores. Sections 2.3 and 2.4 introduce graph-theoretic notation and review standard performance assessment based on recovery of individual edges. In Sections 2.5 and 2.6 we introduce several novel MSSs for assessment of network inference algorithms. We require each MSS to satisfy our list of desiderata; however these scores differ from existing scores by assessing inferred network structure on several (in fact all) scales. For each MSS we discuss associated theoretical and computational issues. Finally Section 2.7 describes computation of p -values for the proposed MSSs.

2.1 Problem Specification

Performance assessment for network inference algorithms may be achieved by comparing estimated networks against known benchmark information. The interpretation of the estimated networks themselves has often been confused in existing literature, with no distinction drawn

between the contrasting notions of significance and effect size. In this Section we therefore formally state our assumptions on the interpretation of both the benchmark network G and the network estimators or estimates \hat{G} .

A1 All networks are directed, unsigned and contain no self-edges.

A network is *signed* if each edge carries an associated $+/-$ symbol. (A1) is widely applicable since an undirected edge may be recast as two directed edges and both signs and self-edges may simply be removed. The challenge of inferring signed networks and more generally the problem of predicting interventional effects requires alternative performance scores that are not dealt with in this contribution, but are surveyed briefly in Sec. 4. The preclusion of self-edges aids presentation but it not required by our methodology.

The form of this benchmark information will influence the choice of performance score and we therefore restrict attention to the most commonly encountered scenario:

A2 The benchmark network G is unweighted.

A network is *weighted* if each edge has an associated weight $w \in \mathbb{R}$. Note that the case of unweighted benchmark networks is widely applicable, since weights may simply be removed if necessary. We will write \mathcal{G}_0 for the space of all directed, unweighted networks that do not contain self-edges and write \mathcal{G} for the corresponding space of directed, weighted networks that do not contain self-edges.

A3 The benchmark network G contains at least one edge and at least one non-edge.

A4 Network estimators \hat{G} are weighted ($\hat{G} \in \mathcal{G}$), with weights having the interpretation that larger values indicate a larger (marginal) probability of the corresponding edge being present in the benchmark network.

In particular we do not consider weights that instead correspond to effect size (see Sec. 4).

A5 In all networks, edges refer to a *direct* dependence of the child on the parent at the level of the vertex set V ; that is, not occurring via any other species in V .

Assumptions (A1-5) are typical for comparative assessment challenges such as DREAM (Marbach *et al.*, 2010; Prill *et al.*, 2010).

2.2 Performance Score Desiderata

Fix a benchmark network G . A *performance score* is defined as function $S : \mathcal{G} \times \mathcal{G}_0 \rightarrow \mathbb{R}$ that accepts an estimated network $\hat{G} \in \mathcal{G}$ and a benchmark network $G \in \mathcal{G}_0$ and returns a real value $S(\hat{G}, G)$ that summarises some aspect of \hat{G} with respect to G . Examples of performance scores are given below. Our approach revolves around certain desiderata that any (i.e. not just multi-scale) performance score S ought to satisfy:

D0 (Interpretability) $S(\hat{G}, G) \in [0, 1]$ for all $\hat{G} \in \mathcal{G}$, $G \in \mathcal{G}_0$, with larger values corresponding to better performance at some specified aspect of network reconstruction.

D1 (Computability) S should be readily computable.

D2 (Objectivity) S should contain no user-specified parameters.

A network estimate $\hat{G} \in \mathcal{G}$ is called *S-optimal* for a benchmark network G if it maximises the performance score $S(\cdot, G)$ over all networks in \mathcal{G} .

D3 (Consistency) The oracle estimator $\hat{G} = G$ is *S-optimal*.

A *rank-preserving* transformation $\{x_i\} \mapsto \{x'_i\}$ of a collection of real values $x_i \in \mathbb{R}$ satisfies $x'_i < x'_j$ whenever $x_i < x_j$.

D4 (Invariance) S should be invariant to rank-preserving transformations of the weights associated with an estimate \hat{G} .

The criteria (D0-3) are important for practical reasons; (D4) is more technical and reflects the fact that we wish to compare estimators whose weights need not belong to the same metric space. i.e. different algorithms may be compared on the same footing, irrespective of the actual interpretation of edge weights (subject to (A4)). Given that much of the popularity of standard classifier scores derives from (D0-4), it is important that any proposed MSS also satisfies the above desiderata. As we will see below, previous studies, such as Banerjee and Jost (2009); Jurman *et al.* (2011a,b); Peters and Bühlmann (2013), do not satisfy (D2,4), precluding their use in objective assessment such as the DREAM Challenges. In Secs. 2.5, 2.6 below we present novel MSS that satisfy each of (D1-4).

2.2.1 Examples of Performance Scores

The local performance scores that have become a standard in the literature, defined in Sec. 2.4 below, are easily shown to satisfy the above desiderata. To date, the challenge of assessing network inference algorithms over multiple scales has received little statistical attention; yet there are several general proposals for quantifying higher order network structure. Notably Jurman *et al.* (2011b) assessed spectral distances for suitability in application to biological networks. The authors recommended use of the Ipsen-Mikhailov distance (IMD) due to its perceived stability and robustness properties, as quantified on randomly generated and experimentally obtained networks (Jurman *et al.*, 2011a). IMD evaluates the difference of the distribution of Laplacian eigenvalues between two networks and therefore could be considered an MSS. There has, to date, been no examination of IMD and related spectral metrics in the context of performance assessment for network inference applications. However IMD and related constructions (e.g. Banerjee and Jost, 2009) fail to satisfy the above desiderata: (i) IMD itself fails to satisfy (D4) since eigenvalues are not invariant to rank-preserving transformations; (ii) the IMD formula contains a user-specified parameter, invalidating (D2).

Peters and Bühlmann (2013) recently introduced “Structural Intervention Distance” (SID; actually a pre-distance) which interprets inferred networks \hat{G}^τ as estimators of an

underlying causal graph G , where both \hat{G}^τ and G must be acyclic (Pearl, 2000). Specifically, SID is the count of pairs of vertices (i, j) for which the estimate \hat{G}^τ incorrectly predicts intervention distributions (in the sense of Pearl (2000)) within the class of distributions that are Markov with respect to G . As such, SID is closely related to the challenge of predicting the effect of unseen interventions. However SID fails to satisfy (D0) due to the requirement that estimated and benchmark networks must be acyclic (i.e. a causal interpretation will not be justified in general) and is therefore unsuitable for this application.

2.3 Notation

Below we introduce notation that will be required to define our proposed scores: Throughout this paper an unweighted network $G \in \mathcal{G}_0$ on vertices $V = \{1, 2, \dots, p\}$ (where $p < \infty$) is treated as a binary matrix $G \in \{0, 1\}^{p \times p}$ with (i, j) th entry denoted $G(i, j)$. When $G(i, j) = 1$ or $G(i, j) = 0$ we say that G contains or does not contain the edge (i, j) , respectively. Write $G(\bullet, j) = \{i \in V : G(i, j) = 1\}$ for the set of parents for vertex $j \in V$. In this paper we do not allow self-edges (A1), so that $G(i, i) = 0$ for all $i \in V$. A path P from i to j in $G \in \{0, 1\}^{p \times p}$ is characterised by a sequence of vertices

$$P = (p_0, p_1, \dots, p_{m-1}, p_m) \in V^{m+1}, \quad p_0 = i, p_m = j \quad (1)$$

such that $G(p_{k-1}, p_k) = 1$ and $p_k \neq i$ for all $k \geq 1$. Note that a path P may contain cycles, but these cycles may not involve p_0 . We say that this path has length $\ell(P) = m$ where $1 \leq m < \infty$. Let $G(i \rightarrow j)$ denote the set of all paths in G from i to j , so that in particular $G(i \rightarrow i) = \emptyset$.

We identify weighted, unsigned networks $\hat{G} \in \mathcal{G}$ with non-negative real-valued matrices $\hat{G} \in [0, 1]^{p \times p}$ and denote the (i, j) th entry by $\hat{G}(i, j)$. Should a network inference procedure produce edge weights in $[0, \infty)$ then by dividing through by the largest weight produces a network with weights in $[0, 1]$. This highlights the importance of property (D4), invariance to rank preserving transformations. It will be required to map \hat{G} into the space of unweighted networks by thresholding the weights at a certain level τ . For $0 \leq \tau \leq 1$ we write \hat{G}^τ for the unweighted network corresponding to a matrix with entries $\hat{G}^\tau(i, j) = \mathbb{I}\{\hat{G}(i, j) \geq \tau\}$.

2.4 Local Scores

In this Section we briefly review local scoring, which has become an established standard in the network inference literature.

Definition: Classification performance scores are defined as functions of confusion matrices, that count the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) calls produced by a classifier. From (A4) it follows that the k largest entries in a network estimate \hat{G} correspond to the k pairs of vertices with largest marginal probabilities of being present as edges in the benchmark network G ; it is therefore reasonable to threshold entries of \hat{G} , say at a level τ , and to consider confusion matrices corresponding to the classifier \hat{G}^τ . In the standard case of *local* estimation, confusion matrices are given by $\text{TP}(\tau) =$

$\sum_{i,j} \hat{G}^\tau(i,j)G(i,j)$, $FP(\tau) = \sum_{i,j} \hat{G}^\tau(i,j)(1-G(i,j))$, $TN(\tau) = \sum_{i,j} (1-\hat{G}^\tau(i,j))(1-G(i,j))$ and $FN(\tau) = \sum_{i,j} (1-\hat{G}^\tau(i,j))G(i,j)$. Based on these quantities, (local) performance scores may be defined. In particular, one widely used score is the area under the receiver operating characteristic (ROC) curve

$$S_{\text{local}}^{\text{ROC}} = \int \text{TPR} \, d\text{FPR} \quad (2)$$

where $\text{TPR} = \text{TP}/(\text{TP} + \text{FN})$ is the *true positive rate* and $\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$ is the *false positive rate*. $S_{\text{local}}^{\text{ROC}}$ has an interpretation as the probability that a randomly selected pair from $\{(i,j) : G(i,j) = 1\}$ is assigned a higher weight $\hat{G}(i,j)$ than a randomly selected pair from $\{(i,j) : G(i,j) = 0\}$ (Fawcett, 2006). As such $S_{\text{local}}^{\text{ROC}}$ takes values in $[0, 1]$ with 1 representing perfect performance and $1/2$ representing performance that is no better than chance. For finitely many test samples, $S_{\text{local}}^{\text{ROC}}$ curves may be estimated by linear interpolation of points $(\text{FPR}(\tau), \text{TPR}(\tau))$ in ROC-space.

Precision-recall (PR) curves are an alternative to ROC curves that are useful in situations where the underlying class distribution is skewed, for example when the number of negative examples greatly exceeds the number of positive examples (Davis and Goadrich, 2006). For biological networks that exhibit sparsity, including gene regulatory networks (Tong *et al.*, 2004) and metabolic networks (Jeong *et al.*, 2000), the number of positive examples (i.e. edges) is frequently smaller than the number of negative examples (i.e. non-edges). In this case performance is summarised by the area under the PR curve

$$S_{\text{local}}^{\text{PR}} = \int \text{PPV} \, d\text{TPR}, \quad (3)$$

where $\text{PPV} = \text{TP}/(\text{TP} + \text{FP})$ is the *positive predictive value*. (In this paper we adopt the convention that $\text{PPV} = 0$ whenever $\text{TP} + \text{FP} = 0$.) $S_{\text{local}}^{\text{PR}}$ also takes values in $[0, 1]$, with larger values representing better performance. For finitely many test samples in PR-space, unlike ROC-space, linear interpolation leads to over-optimistic assessment of performance. The MATLAB code provided in the Supplement follows Goadrich *et al.* (2004) in using nonlinear interpolation to achieve correct, unbiased estimation of area under the PR curve.

Captures: Local scores capture the ability of estimators to recover the exact placement of individual edges. Both $S_{\text{local}}^{\text{ROC}}$ and $S_{\text{local}}^{\text{PR}}$ summarise the ability to correctly infer local topology across a range of thresholds τ , with PR curves prioritising the recovery of positive examples in the “top k edges” (Fawcett, 2006).

Desiderata: (D1; computability), (D2; objectivity) and (D3; consistency) are clearly satisfied. (D4; invariance) is satisfied, since the image of a parametric curve is invariant to any monotone transformation of the parameter (in this case τ).

2.5 Multi-Scale Score 1 (MSS1)

We now introduce the first of our MSSs, which targets ability to infer the connected components of the benchmark network G , through the existence or otherwise of directed paths between vertices.

Definition: In the notation of Sec. 2.3, local performance scores are based on estimation of class labels $Z(i, j) = G(i, j)$ associated with individual edges. A natural generalisation of this approach is to assign labels $Z(i, j) \in \{0, 1\}$ to pairs of vertices $(i, j) \in V \times V$, such that $i \neq j$, based on *descendancy*; that is, based on the presence or otherwise of a directed path from vertex i to vertex j in the benchmark network. i.e. $Z(i, j) = \mathbb{I}\{G(i \rightarrow j) \neq \emptyset\}$. By comparing descendancies in \hat{G}^τ against descendancies in G , we can compute confusion matrices and, by analogy with local scores, we can construct area-under-the-curve scores by allowing the threshold τ to vary. We denote these scores respectively as $S_{\text{MSS1}}^{\text{ROC}}$ and $S_{\text{MSS1}}^{\text{PR}}$. By analogy with local scores, both $S_{\text{MSS1}}^{\text{ROC}}$, $S_{\text{MSS1}}^{\text{PR}}$ take values in $[0, 1]$, and $S_{\text{MSS1}}^{\text{ROC}}$ is characterised as the probability that a randomly selected pair from $\{(i, j) : G(i \rightarrow j) \neq \emptyset\}$ is assigned a higher weight

$$\tau_{i,j} = \max_{P \in \hat{G}^\tau(i \rightarrow j)} \min_{0 \leq i \leq \ell(P)} \hat{G}(p_i, p_{i+1}) \quad (4)$$

than a randomly selected pair from $\{(i, j) : G(i \rightarrow j) = \emptyset\}$.

Captures: MSS1 captures the ability to identify ancestors and descendants of any given vertex. MSS1 scores therefore capture the ability of estimators to recover connected components on all length scales.

Desiderata: (D1; computability) is satisfied due to the well-known Warshall algorithm for finding the transitive closure of a directed, unweighted network (Warshall, 1962), with computational complexity $\mathcal{O}(p^3)$. For the estimated network \hat{G} it is required to know, for each pair (i, j) the value of $\tau_{i,j}$, i.e. the largest value of the threshold τ at which $\hat{G}^\tau(i \rightarrow j) \neq \emptyset$. To compute these quantities we generalised the Warshall algorithm to the case of weighted networks; see Sec. 3.1. (D2-4) are automatically satisfied analogously to local scores, by construction of confusion matrices and area under the curve statistics.

2.6 Multi-Scale Score 2 (MSS2)

Whilst MSS1 scores capture the ability to infer connected components, they do not capture the graph theoretical notion of *differential connectivity* (i.e. the minimum number of edges that must be removed to eliminate all paths between a given pair of vertices). Our second proposed MSS represents an attempt to explicitly prioritise pairs of vertices which are highly connected over those pairs which are weakly connected:

Definition: For each pair $(i, j) \in V \times V$ we will compute an *effect* $0 \leq e_{ij} \leq 1$ that can be thought of as the importance of variable i on the regulation of variable j according to the network G (in a global sense that includes indirect regulation). To achieve this we take inspiration from recent work by Feiglin *et al.* (2012) as well as Morrison *et al.* (2005); Winter *et al.* (2012), who exploit spectral techniques from network theory. Since the effect e_{ij} , which is defined below, includes contributions from all possible paths from i to j in the network \hat{G} , it explicitly captures differential connectivity.

We formally define effects for an arbitrary unweighted network H (which may be either

G or \hat{G}^τ). Specifically the effect e_{ij} of i on j is defined as the sum over paths

$$e_{ij} = \delta_{ij} + \sum_{P \in H(i \rightarrow j)} \prod_{k=1}^{\ell(P)} \frac{1}{|H(\bullet, p_k)|} \quad (5)$$

where δ_{ij} is the Kronecker delta. Effects e_{ij} quantify direct and indirect regulation on all length scales. To illustrate this, notice that Eqn. 5 satisfies the recursive property

$$e_{ij} = \frac{1}{|H(\bullet, j)|} \sum_{k \in H(\bullet, j)} e_{ik} \quad (6)$$

(see Sec. 3.1). Intuitively, Eqn. 20 states that the fraction of j 's behaviour explained by i is related to the combined fractions of j 's parents' behaviour that are explained by i . Rephrasing, in order to explain the behaviour of j it is sufficient to explain the behaviour of each of j 's parents. Moreover, if a parent k of j is an important regulator (in the sense that j has only a small number of parents) then the effect e_{ik} of i on k will contribute significantly to the combined effect e_{ij} . Eqn. 20 is inspired by Page *et al.* (1999) and later Morrison *et al.* (2005); Winter *et al.* (2012), but differs from these works in two important respects: (i) For biological networks it is more intuitive to consider normalisation over incoming edges rather than outgoing edges, since some molecular species may be more influential than others. Mathematically, Page *et al.* (1999) corresponds to replacing $|H(\bullet, j)|$ in Eqn. 20 with $|H(i, \bullet)|$. (ii) Page *et al.* (1999) employed a ‘‘damping factor’’ that imposed a multiplicative penalty on longer paths, with the consequence that effects were readily proved to exist and be well-defined. In contrast our proposal does not include damping on longer paths (see discussion of D2 below) and the theory of Page *et al.* (1999) and others does not directly apply in this setting.

Below we write $\mathbf{e} = \{e_{ij}\}$ for the matrix that collects together all effects for the benchmark network G ; similarly denote by $\hat{\mathbf{e}}^\tau$ the matrix of effects for the thresholded estimator \hat{G}^τ . Any well-behaved measure of similarity between $\hat{\mathbf{e}}^\tau$ and \mathbf{e} may be used to define a MSS. We constructed an analogue of a confusion matrix as TP = $\sum_{i,j} \hat{e}_{ij}^\tau e_{ij}$, FP = $\sum_{i,j} \hat{e}_{ij}^\tau (1 - e_{ij})$, TN = $\sum_{i,j} (1 - \hat{e}_{ij}^\tau) (1 - e_{ij})$, FN = $\sum_{i,j} (1 - \hat{e}_{ij}^\tau) e_{ij}$. Repeating the construction across varying threshold τ , we compute analogues of ROC and PR curves. (Note that, unlike conventional ROC curves, the curves associated with MSS2 need not be monotone; see Supp. Sec. B.1.) Finally scores $S_{\text{MSS2}}^{\text{ROC}}$, $S_{\text{MSS2}}^{\text{PR}}$ are computed as the area under these curves respectively.

Captures: MSS2 is a spectral method, where larger scores indicate that the inferred network \hat{G} better captures the *eigenflows* of the benchmark network G (Lakhina *et al.*, 2004). In general neither $S_{\text{MSS2}}^{\text{ROC}}$ nor $S_{\text{MSS2}}^{\text{PR}}$ need have a unique maximiser (see Sec. 3.1). As such, MSS2 scores do not require precise placement of edges, provided that higher-order topology correctly captures differential connectivity.

Desiderata: Eqn. 5 involves an intractable summation over paths: Nevertheless (D1; computability) is ensured by an efficient iterative algorithm related but non-identical to Page *et al.* (1999), described in Sec. 3.1. In order to ensure (D2; objectivity) we did not include a damping factor that penalised longer paths, since the amount of damping would necessarily

depend on the nature of the data and the scientific context. It is important, therefore, to establish whether effects e_{ij} are mathematically well defined in this objective limit. This paper contributes novel mathematical theory to justify the use of MSS2 scores and prove the correctness of the associated algorithm (see Sec. 3.1). As with MSS1, the remaining desiderata (D3-4) are satisfied by construction.

2.7 Significance Levels

In performance assessment of network inference algorithms we wish to test the null hypothesis $H_0 : \hat{G} \sim \mathcal{M}_0$ for an appropriate null model \mathcal{M}_0 . We will construct a one-sided test based on a performance score S and we reject the null hypothesis H_0 when $S(\hat{G}, G) \in (s^*, 1]$ for an appropriately chosen critical value s^* . In general the distribution of a score $S(R, G)$ under a network-valued random variable R displays a nontrivial dependence on the benchmark network G . We therefore follow Marbach *et al.* (2012) and propose a Monte Carlo approach to compute significance levels, though alternative approaches exist including Scutari and Nagarajan (2013). Specifically, significance of an inferred network \hat{G} under a null model \mathcal{M}_0 is captured by the p -value $p(\hat{G}) := \mathbb{P}(S(R, G) \geq S(\hat{G}, G))$ and given by the strong law of large numbers as

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}(S(R_i, G) \geq S(\hat{G}, G)) \xrightarrow{a.s.} p(\hat{G}) \quad \text{as } n \rightarrow \infty \quad (7)$$

where the R_i are independent samples from \mathcal{M}_0 .

The choice of null model \mathcal{M}_0 is critical to the calculation and interpretation of p -values $p(\hat{G})$. In most biological applications, the null would ideally encompass biological sample preparation, experimental data collection, data preprocessing and network estimation; however in practice it is convenient to define a null model $\mathcal{M}_0 \equiv \mathcal{M}_0(\hat{G})$ conditional upon the inferred network \hat{G} (e.g. Marbach *et al.*, 2010). For brevity, we restrict attention to the following choice: A random network $R = (\pi(V), E(\hat{G}))$ is obtained from \hat{G} by applying a uniformly random permutation π to the vertex labels V ; that is, R is a uniform sample from the space of graph isomorphisms of \hat{G} . This choice has the interpretation that, for deterministic network inference algorithms, the null $\mathcal{M}_0(\hat{G})$ corresponds to randomly permuting the variable labels in the experimental dataset, prior to both data preprocessing and network inference. Note that $\mathcal{M}_0(\hat{G})$ results in a trivial (R_i -independent) hypothesis test when \hat{G} is either empty or complete; we do not consider these degenerate cases in this paper.

In Sec. 3.2.4 we empirically assess the power of this test based on both existing and proposed scores S . MATLAB R2013b code `net_assess`, that was used to compute all of the scores and associated p -values used in this paper, is provided in the Supplement.

3 Results

In Sec. 3.1 below we present theoretical results relating to the proposed scores, demonstrating that MSS and local scores can yield arbitrarily different conclusions in settings where local

topology is recovered very well but higher-order topology is recovered very poorly and vice versa. In Sec. 3.2, in order to empirically assess the statistical power our proposed scores, we appeal to the large corpus of data available from the DREAM community, that represents a large and representative sample of network estimators used by the community.

3.1 Theoretical Results

Whilst distributional results are difficult to obtain, we are able to characterise the behaviour of MSS in both favourable and unfavourable limits. Our first result proves non-uniqueness of S -optimal estimates:

Theorem 1 (Non-uniqueness of S -optimal networks). *For each score S_{MSS1}^{ROC} , S_{MSS1}^{PR} , S_{MSS2}^{ROC} and S_{MSS2}^{PR} , the benchmark network G is always S -optimal (D3). However, for $MSS1$ and $MSS2$, S -optimal estimates are not unique.*

This result contrasts with local performance scores, where S_{local}^{ROC} and S_{local}^{PR} are uniquely maximised by the benchmark network G . Intuitively, Theorem 1 demonstrates that it is possible to add or remove individual edges from a network without changing its higher-order features, such as its connected components.

Our second result proves that the MSS are fundamentally different from existing scores by demonstrating it is possible to simultaneously recover local topology with arbitrary accuracy yet fail to correctly accurately recover the higher-order topology (and vice versa):

Theorem 2 (MSS are distinct from local scores). *(a) For any $\epsilon > 0$ there exist a benchmark network $G \in \mathcal{G}_0$ and an estimate $\hat{G} \in \mathcal{G}$ that satisfy (i) $S_{local}^{ROC}, S_{local}^{PR} > 1 - \epsilon$, (ii) $S_{MSS1}^{ROC}, S_{MSS2}^{ROC} < \frac{1}{2} + \epsilon$, and (iii) $S_{MSS1}^{PR}, S_{MSS2}^{PR} < \epsilon$.*

(b) For any $\epsilon > 0$ there exist a benchmark network $G \in \mathcal{G}_0$ and an estimate $\hat{G} \in \mathcal{G}$ that satisfy (i) $S_{MSS1}^{ROC}, S_{MSS2}^{ROC}, S_{MSS1}^{PR}, S_{MSS2}^{PR} > 1 - \epsilon$, (ii) $S_{local}^{ROC} < \frac{1}{2} + \epsilon$, and (iii) $S_{local}^{PR} < \epsilon$.

An important consequence of this result is that estimators that perform well locally, including DREAM Challenge winners, may not be appropriate for scientific enquiry regarding non-local network topology. Given the increasing use of network inference algorithms in systems biology (e.g. Chuang *et al.*, 2007; Breitkreutz *et al.*, 2012), this highlights the need to study performance of network inference algorithms on multiple length scales.

Our final theoretical contribution is to propose and justify efficient algorithms for computation of both MSS1 and MSS2. For MSS1 we generalise the Warshall algorithm to compute the transitive closure of a directed, weighted graph:

Theorem 3 (Computation of MSS1). *For a fixed weighted network H , let $\tau_{i,j} = \max\{\tau : H^\tau(i \rightarrow j) \neq \emptyset\}$ be the largest value of the threshold parameter τ for which there exists a path from i to j in the network H^τ . Then $\boldsymbol{\tau} = \{\tau_{i,j}\}$ may be computed as follows:*

```

 $\boldsymbol{\tau} = H$ 
for  $k \leftarrow 1 : p$  do

```

```

for  $i \leftarrow 1 : p$  do
  for  $j \leftarrow 1 : p$  do
     $\tau_{i,j} \leftarrow \max(\tau_{i,j}, \min(\tau_{i,k}, \tau_{k,j}))$ 
  end for
end for
end for
return  $\tau$ 

```

For MSS2, following Page *et al.* (1999) we formulate effects e_{ij} as solutions to an eigenvalue problem. However our approach is non-identical to Page *et al.* (1999) (see Sec. 2.6); as a consequence we must prove that (i) effects are mathematically well-defined, and (ii) an appropriately modified version of the algorithm of Page *et al.* (1999) converges to these effects. Below we contribute the relevant theory:

Theorem 4 (Computation of MSS2). *The effects $e_{i,j}$ for an unweighted network H exist, are unique, satisfy Eqn. 20 and may be computed as*

$$\mathbf{e}_i = \lim_{n \rightarrow \infty} \mathbf{v}^{(n)}, \quad \mathbf{v}^{(n)} = \frac{\mathbf{M}^{(i)} \mathbf{v}^{(n-1)}}{\|\mathbf{M}^{(i)} \mathbf{v}^{(n-1)}\|}, \quad v_k^{(0)} = \delta_{ik} \quad (8)$$

where \mathbf{e}_i is the i th row of \mathbf{e} and $\mathbf{M}^{(i)}$ is defined by (i) $M_{im}^{(i)} = \delta_{im}$, (ii) $M_{km}^{(i)} = \frac{1}{|H(\bullet, k)|}$ for all $m \in H(\bullet, k)$ and $H(i \rightarrow k) \neq \emptyset$, (iii) $M_{km}^{(i)} = 0$ otherwise.

Theorem 4 facilitates highly efficient numerical computation of the effect matrices \mathbf{e} required for MSS2. In practice Eqn. 8 requires convergence diagnostics; for experiments in this paper we used the stopping rule $\|\mathbf{v}^{(n)} - \mathbf{v}^{(n-1)}\|_1 < 0.01 \times p$.

3.2 Empirical Results

For an objective comparison covering a wide range of network inference methodologies, we exploited the results of the DREAM5 challenge, where 36 methodologies were blind-tested against both simulated and real transcription factor data (Marbach *et al.*, 2012). The network inference challenge included (i) *in silico* data generated using GeneNetWeaver (Marbach *et al.*, 2009), (ii) experimentally validated interactions from a curated database for *E. coli*, and (iii) transcription-factor binding data and evolutionarily conserved binding motifs for *S. Cerevisiae*. Data were provided on the genomic scale, involving (i) 1643, (ii) 4297 and (iii) 5667 genes respectively. For each of these systems, benchmark information was provided in the form of targets for a subset of genes. For assessment purposes we restricted attention to predictions made regarding these subnetworks only, satisfying (A1-3), consisting of (i) 195, (ii) 334 and (iii) 333 genes respectively. The DREAM5 data are well-suited for this investigation, representing a wide variety of network reconstruction algorithms contributed by the community and containing sufficiently many samples to provide a definitive assessment of statistical power.

In each challenge, participants were required to provide a list of inferred edges (i, j) , together with associated weights $0 \leq \hat{G}(i, j) \leq 1$. Such edges were directed, unsigned and excluded self-edges, satisfying (A1). Participants were asked to “infer the structure of the underlying transcriptional regulatory networks”, so that edges were ranked in terms of their probability of existence, rather than their inferred effect size (A4). For DREAM5, (A5; direct dependence) appears to be implicit. The algorithms used to obtain edge lists and weights varied considerably, being classified according to their statistical formulation as either *Bayesian*, *Correlation*, *Meta*, *Mutual Information* (MI), *Regression* or *Other*. For each methodology we obtained performance scores for datasets (i-iii) using the `net_assess` package (Supp. Figs. 3, 4, 5). We also investigated the Overall Score reported by Marbach *et al.* (2012), that summarises performance by combining local AUROC and AUPR scores across all 3 datasets (further details in Supp. Sec. B.3).

In Section 3.2.1 below we examine the characteristics of our proposed scores in an aggregate sense based on the full DREAM5 data. Section 3.2.2 then considers estimator-specific score profiles in order to understand the multi-scale properties of different network inference algorithms. In particular Section 3.2.3 focusses on a “wisdom of crowds” approach that has previously demonstrated strong local performance in this setting. Finally Section 3.2.4 quantifies the statistical power of our proposed scores in the context of hypothesis testing.

3.2.1 Aggregate Results

Local and MSS scores may be arbitrarily different in principle (Theorem 2), however in practice these scores may be highly correlated. To investigate this we produced scatter plots between the existing and proposed performance scores over all 36 estimators and all 3 datasets (a total of $n = 108$ samples; Fig. 1, Supp. Fig. 6). Our findings, based on Spearman correlation (Table 1), showed that all scores were significantly correlated under AUPR on the *in silico* and *E. Coli* datasets ($p \leq 0.01$), but not on the *S. Cerevisiae* dataset. Indeed results for the *S. Cerevisiae* dataset were less impressive for all methods, under all performance scores, in line with the general poor performance on this dataset reported by Marbach *et al.* (2012). On the combined data, both MSS1 and MSS2 was significantly correlated with the local score under AUROC ($\rho = 0.84$, $\rho = 0.57$; $p < 10^{-10}$ in each case) suggesting that accurate reconstruction of local topology is associated with accurate reconstruction of higher-order topology. On the other hand, on all 3 datasets and both AUROC and AUPR statistics, MSS2 was less highly correlated with the local score compared to MSS1 (e.g. on the combined data and AUPR, $\rho = 0.82$ versus $\rho = 0.29$). This suggests that estimated local topology is more predictive of descendancy relationships than predictive of differential connectivity and spectral features. In many cases MSS1 and MSS2 were significantly correlated; this was expected since both estimators target multiple scales by design.

3.2.2 Individual Participant Performance

Network inference algorithms are predominantly employed in contexts where there is considerable uncertainty regarding the data-generating network topology. In such settings attention

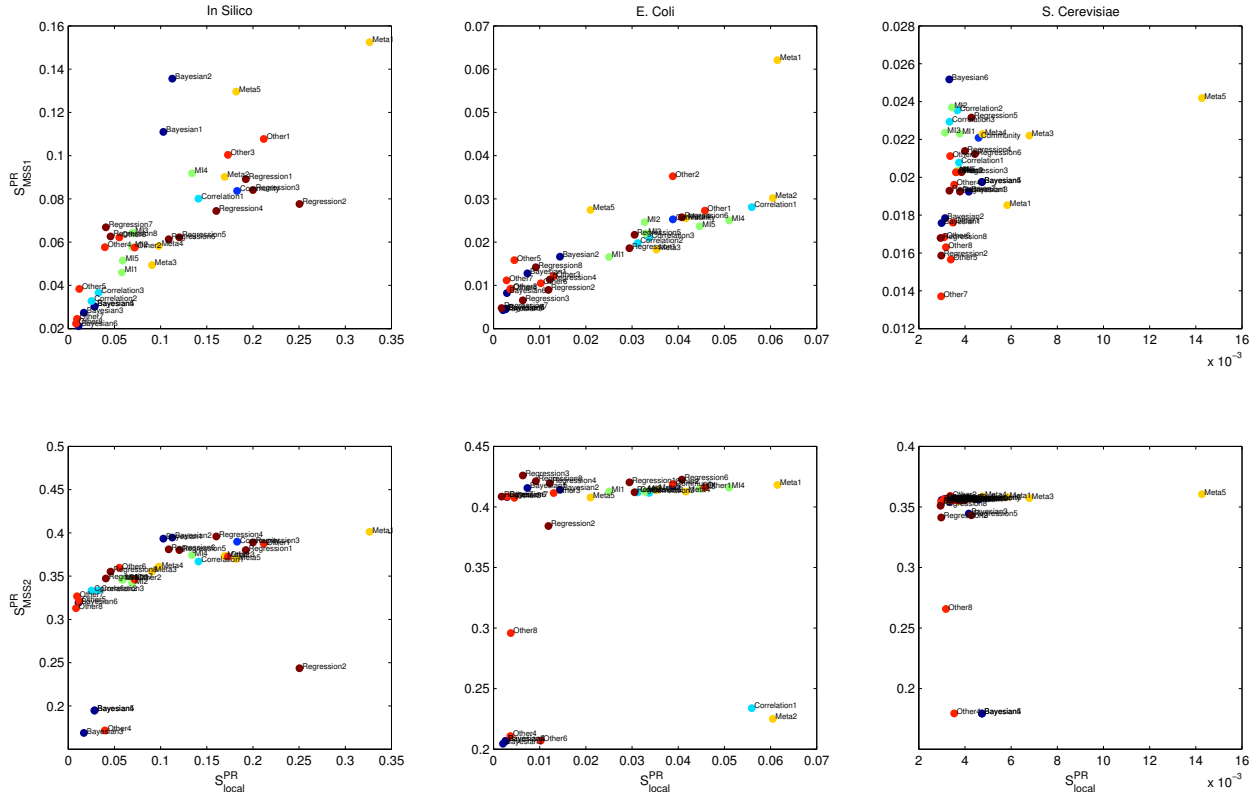


Figure 1: DREAM5 network inference challenge data. [The 36 DREAM5 methodologies were assessed using both local and multi-scale performance scores. Participants were grouped according to their statistical formulation as either *Bayesian*, *Correlation*, *Meta*, *Mutual Information* (MI), *Regression* or *Other*. *Community* represents a crowd-sourced network estimator proposed by Marbach *et al.* (2012). Left to right: *In silico* dataset, *E. Coli* dataset, *S. Cerevisiae* dataset. Here we show results based on area under the precision recall curve.]

is often restricted to the “most significant” aspects of inferred topology; we therefore similarly restrict attention to individual estimator performance as quantified by scores based on AUPR, reserving AUROC results for Supp. Sec. B.4.

Under MSS1 the best performing methods included *Bayesian 1,2,6* and *Meta 1,5* (Fig. 1). For MSS2 many methods attained similarly high scores; these included *Bayesian 1,2,6* and *Meta 1,5*. Interestingly these 5 best performing algorithms did not rank highly according to the (local) Overall Score of Marbach *et al.* (2010) (placing 8th, 24th, 20th, 22nd and 34th respectively out of 36). Similarly *Regression 2*, which almost maximised $S^{\text{PR}}_{\text{local}}$ on the *in silico* dataset, was among algorithms with the lowest $S^{\text{PR}}_{\text{MSS2}}$. This appears to be a real example of the conclusion of Theorem 2, where *Regression 2* recovers several individual edges yet fails faithfully reconstruct higher-order topology, including connected components. Supp. Fig. 7 displays inferred and benchmark topology for *Regression 2*; it is visually clear that the algorithm fails to distinguish between different connected components in the true data-generating network. Conversely, estimators such as *Bayesian 1,2* simultaneously

	Local	Marbach <i>et al.</i>	MSS1	MSS2	Local	Marbach <i>et al.</i>	MSS1	MSS2	
<i>In Silico</i> ($n = 36$)		0.60	0.80	0.64		0.78	0.87	0.77	Local
	$< 10^{-4}$		0.53	0.78	$< 10^{-10}$		0.53	0.52	Marbach <i>et al.</i>
	$< 10^{-10}$	$< 10^{-3}$		0.51	$< 10^{-20}$	$< 10^{-3}$		0.82	MSS1
	$< 10^{-4}$	$< 10^{-10}$	$< 10^{-2}$		$< 10^{-10}$	$< 10^{-2}$	$< 10^{-10}$		MSS2
<i>E. Coli</i> ($n = 36$)		0.42	0.92	0.13		0.58	0.92	0.44	Local
	0.01		0.27	0.46	$< 10^{-3}$		0.50	0.54	Marbach <i>et al.</i>
	$< 10^{-20}$	0.11		0.05	$< 10^{-20}$	$< 10^{-2}$		0.42	MSS1
	0.46	$< 10^{-2}$	0.77		$< 10^{-2}$	$< 10^{-3}$	0.01		MSS2
<i>S. Cerevisiae</i> ($n = 36$)		0.26	0.70	0.17		0.29	0.44	0.18	Local
	0.13		0.13	0.45	0.09		0.22	0.23	Marbach <i>et al.</i>
	$< 10^{-5}$	0.46		-0.06	$< 10^{-2}$	0.19		0.32	MSS1
	0.31	$< 10^{-2}$	0.74		0.29	0.17	0.06		MSS2
Combined ($n = 108$)		0.29	0.84	0.57		0.37	0.82	0.29	Local
	$< 10^{-2}$		0.28	0.34	$< 10^{-4}$		0.27	0.36	Marbach <i>et al.</i>
	$< 10^{-20}$	$< 10^{-2}$		0.40	$< 10^{-20}$	$< 10^{-2}$		0.04	MSS1
	$< 10^{-10}$	$< 10^{-3}$	$< 10^{-4}$		$< 10^{-2}$	$< 10^{-3}$	0.69		MSS2
	AUROC				AUPR				

Table 1: Spearman correlation between performance scores (white) and associated p -values (grey); DREAM5 network inference challenge data. [Performance scores include area under the receiver operating characteristic (AUROC) and precision recall (AUPR) curves, based on both local scores and the proposed Multi-Scale Scores (MSS). The Overall Score of Marbach *et al.* (2012) combines both local AUROC and AUPR scores as described in Supp. Sec. B.3.]

achieve strong multi-scale performance and unimpressive local performance (Fig. 1). It is well understood that Bayesian estimators are well suited to recovery of a coherent joint graphical model, compared to estimators that decompose network inference into independent neighbourhood selection problems (Marbach *et al.*, 2010). These results suggest that this intuition is manifest in the multi-scale performance of Bayesian network inference algorithms as quantified by both S_{MSS1}^{PR} and S_{MSS2}^{PR} .

3.2.3 Community Performance

A key finding of Marbach *et al.* (2012) was that the *Community* network, a principled aggregation of the DREAM5 participants’ predictions, was able to maximise the Overall Score for local performance (the so-called “wisdom of crowds” phenomenon). Interestingly, we found that this *Community* network also performed well over multiple length scales (Supp. Figs. 3,4,5). In fact the *Community* network was almost S_{MSS2}^{ROC} and S_{MSS2}^{PR} -optimal over all 36 methodologies on all 3 datasets and was ranked highly under S_{MSS1}^{ROC} in all experiments. These results suggest that aggregation of estimators that perform well locally (e.g. *Regression 2*) and well on higher-order topology (e.g. *Bayesian 2*) may be a successful strategy to achieve strong performance over multiple length scales.

3.2.4 Significance Testing

A common fault of complex test statistics is that they lack power relative to simpler statistics (e.g. Simon and Tibshirani (2012)). In order to explore the relative statistical power of existing and proposed performance scores S in hypothesis testing, we computed Monte Carlo p -values for each network estimate \hat{G} in the DREAM5 challenge, taking each score in turn as a test function (see Sec. 2.7). Under the assumption that all DREAM5 participants were capable, in principle, of recovering some aspects of structural information (i.e. estimators were not simply random samples from \mathcal{M}_0), we may estimate statistical power by treating the DREAM5 data as cases where H_0 should be rejected. Results (Fig. 2) showed that, over all 36 participants and all 3 datasets, $S_{\text{local}}^{\text{PR}}$ assigned a significance level $p < 0.05$ to 88% of the networks, $S_{\text{MSS1}}^{\text{PR}}$ achieved 73% whilst $S_{\text{MSS2}}^{\text{PR}}$ achieved 71% at this threshold. Varying the threshold $1 > p > 0.01$ we see that $S_{\text{local}}^{\text{PR}}$ is roughly 25% more powerful compared to both $S_{\text{MSS1}}^{\text{PR}}$ and $S_{\text{MSS2}}^{\text{PR}}$. Given that challenge participants likely developed methodology to optimise local performance scores rather than multi-scale scores, these findings support the conclusion that the multi-scale scores $S_{\text{MSS1}}^{\text{PR}}$, $S_{\text{MSS2}}^{\text{PR}}$ offer comparable power relative to $S_{\text{local}}^{\text{PR}}$. From the perspective of methodology, these results show that most DREAM5 algorithms are better suited to identifying local rather than higher-order topology. Results for ROC-based scores demonstrated a more considerable decrease in statistical power between local and multi-scale test statistics; it may therefore be prudent to restrict attention to multi-scale results based on PR-scores.

4 Discussion

Network inference algorithms are increasingly used to facilitate diverse scientific goals, including prediction of single interactions with high precision, identifying motifs such as cliques or clusters, and uncovering global topology such as connected components. Yet performance assessment of these algorithms does not currently distinguish between these contrasting goals; we do not know which algorithms are most suited to which tasks. Widely used local scores based on classifier analysis capture the ability of an algorithm to exactly recover a benchmark network at the local level. In this paper we proposed novel multi-scale scores (MSSs) that instead capture ability to infer connectivity patterns simultaneously over all length scales. Unlike previous multi-scale proposals, our MSS satisfy desiderata that have contributed to the popularity of local performance scores, including invariance to rank-preserving transformation. Theoretical results demonstrated the potential for local and multi-scale analyses to draw arbitrarily different conclusions. Empirical results, based on a comprehensive analysis of the DREAM5 challenge data, confirmed that whilst local and multi-scale performance were correlated, in several cases an estimator recovered local topology well but recovered higher-order topology poorly and *vice versa*. These results highlight the importance of performance assessment on multiple length scales, since apparently promising methodologies may be highly unsuitable for inference of non-local network topology. Notably the *Community* network proposed by Marbach *et al.* (2012) enjoyed both good local and good multi-scale

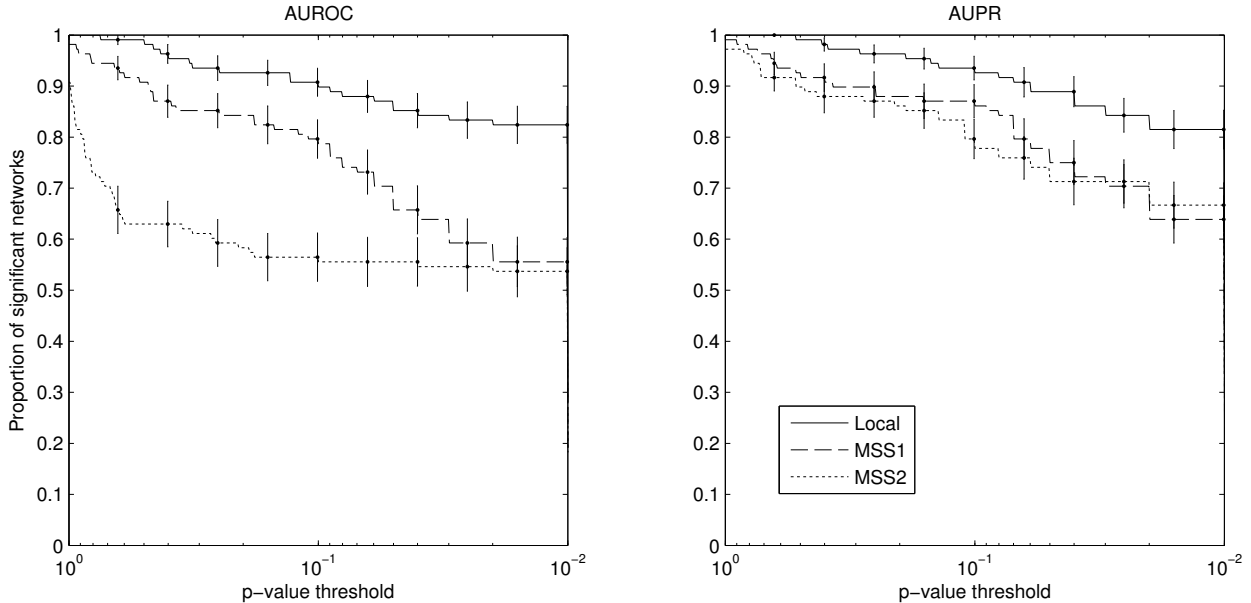


Figure 2: Significance testing; empirical study of statistical power. [Here we considered all 36 DREAM5 methodologies and all 3 datasets, computing Monte Carlo p -values for each of the area under the receiver operating characteristic (AUROC) and precision-recall (AUPR) curves, as defined by both local and multi-scale scores. Error bars represent the 68% confidence region, computed via bootstrap resampling.]

performance. Finally we demonstrated that one proposed score (S_{MSS2}^{PR}) possesses comparable statistical power with respect to local scores.

Our proposals extend scores based on individual edges by considering paths of length > 1 . Multi-scale analysis of networks is now an established research field; indeed techniques based on Page *et al.* (1999) have recently come to light in both gene regulatory network (Morrison *et al.*, 2005; Winter *et al.*, 2012) and protein signalling network (Johannes *et al.*, 2010; Feiglin *et al.*, 2012) analyses. This work differs to these within-network studies by focussing on the challenge of performance assessment. The statistical comparison of two unweighted networks has recently been considered by Städler and Mukherjee (2013); Yates and Mukhopadhyay (2013). In related work, Milenković *et al.* (2009) constructed an *ad-hoc* null distribution over residue interaction graphs based on degree-distribution, clustering coefficient, clustering spectrum, average network diameter and spectrum of shortest path lengths. However these are *mean field* statistics, which do not respect vertex labels, making their use inappropriate for many scientific applications.

Existing and proposed scores capture complementary aspects of performance, so that it is misleading to speak of a universally “best” methodology. MSS was specifically designed to capture aspects of higher-order network reconstruction that have been relatively neglected by the community, however in general the existing and proposed scores together still fail to capture many important aspects of estimator performance. In that respect, this paper rep-

resents a small step towards more comprehensive characterisation of estimator performance. MSS1 required a factor of $\mathcal{O}(p)$ more computation than existing local scores, with MSS2 requiring a factor of $\mathcal{O}(p^2)$. (Computational times are presented in Supp. Fig. 8.) Given that in practice it may be desirable to restrict attention to a single measure of multi-scale performance, and given that $S_{\text{MSS1}}^{\text{PR}}$ and $S_{\text{MSS2}}^{\text{PR}}$ provide similar statistical power, it may be preferable to favour MSS1 over MSS2.

Whilst multi-scale analysis has the potential to complement local analyses and deepen our understanding of the applicability and limitations of network inference algorithms, there are two commonly encountered settings where multi-scale analyses may actually be preferred over local analyses:

1. **Time series data:** Iwasaki and Simon (1994); Dash (2003) and others have noticed that, in the setting of causal graphical models and time series data, the “true” causal graph G depends crucially on the time scale at which the process is described. To see this, notice that the simple multivariate process defined by the causal graph G , equal to $\dots \rightarrow X_i \rightarrow X_{i+1} \rightarrow \dots$ and distributions $X_{i+1}(t+1) = X_i(t) + \epsilon_{i+1}(t+1)$ with $\epsilon_i(t)$ independent $\mathcal{N}(0, \sigma^2)$ random variables, satisfies $X_{i+2}(t+2) = X_i(t) + \tilde{\epsilon}_{i+2}(t+2)$ with $\tilde{\epsilon}$ independent $\mathcal{N}(0, \tilde{\sigma}^2)$ random variables. Thus any consistent estimator of a linear Gaussian vector autoregressive process will, based on data from even time points only, infer the causal graph $\hat{G} = G^2$ equal to $\dots \rightarrow X_i \rightarrow X_{i+2} \rightarrow \dots$ i.e. the apparently natural estimator demonstrates the worst possible performance under $S_{\text{local}}^{\text{ROC/PR}}$. This is undesirable since it remains the case that all descendancy relationships in \hat{G} are all contained in G . However \hat{G} is readily seen to be $S_{\text{MSS1/MSS2}}^{\text{PR}}$ -optimal, so that the proposed scores are robust to the problem of inappropriate sampling frequency. Real-world examples of this problem arise in the (frequently encountered) settings where no natural time scale is available for experimental design, or the sampling frequency is limited by resource constraints.
2. **Interventional experiments:** In experimental settings, network topology is frequently inferred through targeted interventions (e.g. knock-downs, knock-outs, small molecule inhibitors, monoclonal antibodies etc.). Differential analysis with respect to a control sample in this case does not reveal the direct children of the interventional target, but rather specifies the descendants of that target. Thus benchmark information is naturally available on descendants, whereas additional experimental work is required to uncover direct edges. In these settings performance assessment using descendancy (i.e. MSS1) rather than using individual edges facilitates a reduction in experimental cost.

This paper focused on the problem of inference for network structure, but an alternative viewpoint is to assess networks by their predictive power (Prill *et al.*, 2011). In this setting (A4) is replaced by an assumption that estimator weights correspond to effect size (and may be signed). For example these weights may be average causal effects (ACE; Pearl, 2000; Maathuis *et al.*, 2010). However this setting is less common as many popular network inference methodologies do not entail a predictive statistical model. Moreover, prediction

is often computationally intensive, whereas assessment of the inferred topology is relatively cheap.

5 Conclusion

This paper lays statistical and conceptual foundations for the analysis of network inference algorithms on multiple scales. We restricted attention to the important problem of comparing between inference procedures, as in the DREAM challenges, and proposed novel multi-scale scores (MSSs) for this setting. An empirical study based on the DREAM5 data demonstrated that multi-scale analysis provides additional insight into the character and capability of network inference algorithms and suggested that a crowd-source approach to inference may offer improved reconstruction of higher-order topology. In this paper we focussed on connectivity, but multi-scale scores may be designed to capture ability to infer specific motifs such as cliques, or particular feedback circuits, for example. This work is exploratory and should not be used to form conclusions regarding the performance of specific teams in the DREAM5 challenge.

Acknowledgement

The authors are grateful to Gustavo Stolovitzky and Sach Mukherjee for feedback on an earlier draft of this manuscript. This work was supported by the Centre for Research in Statistical Methodology grant (EPSRC EP/D002060/1). CJO was involved in designing the DREAM8 Challenge. CJO, RA, SEFS were not involved in the DREAM5 Challenge.

A Proofs

Proof of Theorem 1. S -optimality of the benchmark network G follows immediately from the definition of each score. To prove nonuniqueness we begin by considering MSS1. For this, take G to be the network $A \rightarrow B \rightarrow C \rightarrow A$ and take G' to be the network $A \leftarrow B \leftarrow C \leftarrow A$; then both networks entail the same descendancy relationships. It follows from (D3; consistency) that both G and G' maximise $S_{\text{MSS1}}^{\text{ROC}}$ and $S_{\text{MSS1}}^{\text{PR}}$, demonstrating that the maximiser will be non-unique in general. The argument for MSS2 is analogous, using the same pair of networks. \square

Proof of Theorem 2. (a) Given $\epsilon > 0$. We proceed by constructing a sequence of pairs (G, \hat{G}) indexed by p , the number of vertices, such that the scores associated with \hat{G} end simultaneously to the required limits as $p \rightarrow \infty$. Define a data-generating network G on p vertices by the edge set

$$E(G) = \{(1, i) : 2 \leq i \leq p - 1\} \cup \{(i, p) : 2 \leq i \leq p - 1\} \quad (9)$$

and consider an unweighted network \hat{G} with topology $E(\hat{G}) = E(G) \cup \{(p, 1)\}$. i.e. \hat{G} differs to G in the addition of a single edge $(p, 1)$. The ROC curve corresponding to \hat{G} is defined by the three points

$$\{(0, 0), (\text{FPR}, \text{TPR}), (1, 1)\} \quad (10)$$

so that by linear interpolation the area under the curve is $\frac{1}{2}(1 + \text{TPR} - \text{FPR})$. The PR curve is similarly defined by

$$\left\{ (0, 0), (\text{TPR}, \text{PPV}), \left(1, \frac{T}{p(p-1)}\right) \right\}. \quad (11)$$

Here T is the number of true examples; for local scores this is $E(G)$, for MSS1 this is $\#\{(i, j) : G(i \rightarrow j) \neq \emptyset, i \neq j\}$ and for MSS2 this is $\sum_{i,j} e_{i,j}$. To compute the area under this curve we must use nonlinear interpolation (Goadrich *et al.*, 2004). Specifically, interpolation between the two points A, B with true/false positive values $(\text{TP}_A, \text{FP}_A), (\text{TP}_B, \text{FP}_B)$ respectively we create new points for each of $\text{TP}_A + 1, \text{TP}_A + 2, \dots, \text{TP}_B - 1$, increasing the false positives for each new point by $(\text{FP}_B - \text{FP}_A)/(\text{TP}_B - \text{TP}_A)$. By direct calculation we have that the area under this interpolated curve is

$$\frac{\text{TP}^2}{T(\text{TP} + \text{FP})} + \frac{1}{T} \sum_{j=1}^{T-\text{TP}} \frac{(\text{TP} + j)(T - \text{TP})}{(\text{TP} + \text{FP})(T - \text{TP}) + j[p(p-1) - \text{TP} - \text{FP}]}. \quad (12)$$

Fix $\epsilon > 0$. For (i), since \hat{G} differs to G by just a single edge we have $\text{TP} = 2(p-2)$, $\text{FP} = 1$, $T = 2(p-2)$, $\text{TPR} = 1$, $\text{FPR} = \frac{1}{(p-1)(p-2)}$, $\text{PPV} = \frac{2p-4}{2p-3}$. It is then easily checked that

$$S_{\text{local}}^{\text{ROC}} = \frac{1}{2} \left(1 + 1 - \frac{1}{(p-1)(p-2)} \right) \rightarrow 1 \quad \text{as } p \rightarrow \infty, \quad (13)$$

$$S_{\text{local}}^{\text{PR}} = \frac{[2(p-2)]^2}{[2(p-2)][2(p-2) - 1]} \rightarrow 1 \quad \text{as } p \rightarrow \infty. \quad (14)$$

Thus $\exists P_1 \in \mathbb{N} : \forall p \geq P_1, S_{\text{local}}^{\text{ROC}} > 1 - \epsilon, S_{\text{local}}^{\text{PR}} > 1 - \epsilon$.

For (ii-iii) and MSS1 note that $\hat{G}(i \rightarrow j) \neq \emptyset$ for all $i \neq j$, so that \hat{G} predicts every possible descendency. Thus $\text{TP} = 3(p-2)$, $\text{FP} = p^2 - 4p + 1$, $T = 3(p-2)$, $\text{TPR} = 1$, $\text{FPR} = 1$, $\text{PPV} = \frac{3(p-2)}{p(p-1)}$ and

$$S_{\text{MSS1}}^{\text{ROC}} = \frac{1}{2} (1 + 1 - 1) \rightarrow \frac{1}{2} \quad \text{as } p \rightarrow \infty, \quad (15)$$

$$S_{\text{MSS1}}^{\text{PR}} = \frac{[3(p-2)]^2}{[3(p-2)][3(p-2) + (p^2 - 4p + 1)]} \rightarrow 0 \quad \text{as } p \rightarrow \infty. \quad (16)$$

Thus $\exists P_2 \in \mathbb{N} : \forall p \geq P_2, S_{\text{MSS1}}^{\text{ROC}} < \frac{1}{2} + \epsilon, S_{\text{MSS1}}^{\text{PR}} < \epsilon$. For MSS2 note that the estimated effects

are $\hat{e} = \mathbf{1}_{p \times p}$ whereas the benchmark effects are

$$\mathbf{e} = \begin{bmatrix} 1 & 1 & \dots & 1 & 1 \\ & 1 & & & (p-2)^{-1} \\ & & \ddots & & \vdots \\ & & & & (p-2)^{-1} \\ & & & & 1 \end{bmatrix} \quad (17)$$

leading to $TP = 2p$, $FP = p(p-2)$, $T = 2p$, $TPR = 1$, $FPR = 1$, $PPV = \frac{2}{p}$ and

$$S_{\text{MSS2}}^{\text{ROC}} = \frac{1}{2} (1 + 1 - 1) \rightarrow \frac{1}{2} \quad \text{as } p \rightarrow \infty, \quad (18)$$

$$S_{\text{MSS2}}^{\text{PR}} = \frac{[2p]^2}{[2p][2p + p(p-2)]} \rightarrow 0 \quad \text{as } p \rightarrow \infty. \quad (19)$$

Thus $\exists P_3 \in \mathbb{N} : \forall p \geq P_3, S_{\text{MSS2}}^{\text{ROC}} < \frac{1}{2} + \epsilon, S_{\text{MSS2}}^{\text{PR}} < \epsilon$.

Taking $p > \max\{P_1, P_2, P_3\}$, we have shown that (G, \hat{G}) satisfy the conclusions of the theorem.

(b) The converse result is proved similarly, taking G to be the cycle $1 \rightarrow 2 \rightarrow \dots \rightarrow p \rightarrow 1$ and \hat{G} to be the reverse $1 \leftarrow 2 \leftarrow \dots \leftarrow p \leftarrow 1$. \square

Proof of Theorem 3. Let $P = (p_0, \dots, p_m)$ be a path from p_0 to p_m . Define the intermediate vertices in the path to be the vertices p_1, \dots, p_{m-1} . For each pair of vertices i and j we wish to calculate $\tau_{i,j}$, the largest threshold τ for which i and j are connected by a path in H^τ .

For $k = 0, \dots, p$, let $\tau_{i,j}^{(k)}$ represent the largest value of the threshold τ for which H^τ contains a path from i to j with intermediate vertices in the set $\{1, \dots, k\}$. From this definition, we have $\tau_{i,j}^{(0)} = H$, the weighted adjacency matrix and $\tau_{i,j} = \tau_{i,j}^{(p)}$. Next we show that for $k = 1, \dots, p$,

$$\tau_{i,j}^{(k)} = \max\{\tau_{i,j}^{(k-1)}, \min\{\tau_{i,k}^{(k-1)}, \tau_{k,j}^{(k-1)}\}\}. \quad (20)$$

To see this, note that $\tau_{i,j}^{(k)} = \tau_{i,j}^{(k-1)}$ unless there is a more strongly connected path which goes through k . Such a path must first connect i to k , and then connect k to j , with both paths involving intermediate vertices in $\{1, \dots, k-1\}$. This connection is broken in H^τ once the threshold τ is raised above the minimum of these two connections, that is when $\tau > \min\{\tau_{i,k}^{(k-1)}, \tau_{k,j}^{(k-1)}\}$.

The conclusion of the Theorem follows by iterative application of Eqn. 20. \square

Remarks on the proof of Theorem 3: To make the algorithm more memory efficient, we note that it is not necessary to store more than $\mathcal{O}(p^2)$ values for τ at any point. This stems from the fact that each iteration is calculated only from the results of the previous one, reducing the necessary storage to $2p^2$. This can be improved further by noting that if $i = k$ or $j = k$ then k cannot be an intermediate vertex in a path connecting i to j and hence $\tau_{i,j}^{(k)} = \tau_{i,j}^{(k-1)}$. These $2k - 1$ index values are the only elements needed to calculate the k th iteration, and

they themselves do not need to be updated. Hence, the remaining elements in the matrix are needed only for their own update and can be safely over-written. A corollary to this is that the i and j loops in the algorithm can be completed in any order or even in parallel. Note that this algorithm works whether or not the weight matrix H is allowed to contain loops. Irrespective of whether loops are permitted or not, $\tau_{i,i}$ represents the largest value of the edge threshold for which node i is contained in a cycle.

For Theorem 4 we require a series of Lemmas:

Lemma 1. *The effects*

$$e_{ij} = \delta_{ij} + \sum_{P \in \mathcal{P}(i,j)} \prod_{k=1}^{\ell(P)} \frac{1}{|H(\bullet, p_k)|} \quad (21)$$

are well-defined and satisfy $0 \leq e_{ij} \leq 1$.

Proof. We show the sum in Eqn. 21 converges absolutely. Since all terms are non-negative, it is sufficient to prove that, for each $i \neq j$ such that $\mathcal{P}(i, j) \neq \emptyset$, the sum is bounded above. Consider the partial sums

$$e_{ij}^{(n)} = \sum_{P \in \mathcal{P}(i,j): \ell(P) \leq n} \prod_{k=1}^{\ell(P)} \frac{1}{|H(\bullet, p_k)|}. \quad (22)$$

If Eqn. 21 diverges, so must $\lim_{n \rightarrow \infty} e_{ij}^{(n)}$ since all paths are of finite length, so it is sufficient to prove $e_{ij}^{(n)}$ is bounded above by one as $n \rightarrow \infty$. We proceed inductively, with base case

$$e_{ij}^{(1)} = \left\{ \begin{array}{ll} \frac{1}{|H(\bullet, j)|} & : i \in H(\bullet, j) \\ 0 & : \text{otherwise} \end{array} \right\} \leq 1. \quad (23)$$

Suppose $e_{ij}^{(n-1)} \leq 1$. Then for $i \notin H(\bullet, j)$ we have

$$e_{ij}^{(n)} = \sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} e_{ik}^{(n-1)} \quad (24)$$

since a path P of length m can be decomposed into a path from $p_0 = i$ to $p_{m-1} = k$ and an edge from $p_{m-1} = k$ to $p_m = j$. Therefore $e_{ij}^{(n)} \leq \sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} = 1$ as required. The case where $i \in H(\bullet, j)$ is similar, leading to

$$e_{ij}^{(n)} = \sum_{k \in H(\bullet, j) \setminus \{i\}} \frac{1}{|H(\bullet, j)|} e_{ik}^{(n-1)} \quad (25)$$

from which we again conclude $e_{ij}^{(n)} \leq \sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} = 1$. Therefore $e_{ij}^{(n)} \leq 1$ for all n by induction. \square

Lemma 2. *Effects, as defined in Eqn. 21, satisfy (i) $e_{ij} = \delta_{ij}$ for all i, j s.t. $\mathcal{P}(i, j) = \emptyset$, (ii) $e_{ij} = \sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} e_{ik}$ for all i, j s.t. $\mathcal{P}(i, j) \neq \emptyset$.*

Proof of Lemma 2. The first statement follows immediately from Eqn. 21. For the second statement, consider two cases: Firstly, for $i \neq j$ such that $i \notin H(\bullet, j)$ we have

$$\sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} e_{ik} = \sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} \sum_{P \in \mathcal{P}(i, k)} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} \quad (26)$$

$$= \sum_{P \in \mathcal{P}(i, j)} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} = e_{ij}. \quad (27)$$

Note that absolute convergence (Lemma 1) ensures the manipulation of infinite sums is valid. The proof for $i \in H(\bullet, j)$ is similar but requires slightly more care:

$$\sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} e_{ik} = \sum_{k \in H(\bullet, j)} \frac{1}{|H(\bullet, j)|} \left\{ \delta_{ik} + \sum_{P \in \mathcal{P}(i, k)} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} \right\} \quad (28)$$

$$= \frac{1}{|H(\bullet, j)|} + \sum_{k \neq i} \frac{1}{|H(\bullet, j)|} \sum_{P \in \mathcal{P}(i, k)} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} \quad (29)$$

$$= \sum_{\substack{P \in \mathcal{P}(i, j) \\ p_{\ell(P)-1} = i}} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} + \sum_{\substack{P \in \mathcal{P}(i, j) \\ p_{\ell(P)-1} \neq i}} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} \quad (30)$$

$$= \sum_{P \in \mathcal{P}(i, j)} \prod_{l=1}^{\ell(P)} \frac{1}{|H(\bullet, p_l)|} = e_{ij}. \quad (31)$$

□

Lemma 3. *Define a matrix $\mathbf{M}^{(i)}$ by (i) $M_{im}^{(i)} = \delta_{im}$, (ii) for $\mathcal{P}(i, k) \neq \emptyset$ and $m \in H(\bullet, k)$, set $M_{km}^{(i)} = \frac{1}{|H(\bullet, k)|}$, (iii) $M_{km}^{(i)} = 0$ otherwise. Then the conclusions of Lemma 2 are equivalent to the following: For each i , the vector \mathbf{v} where $v_k = e_{ik}$ satisfies $\mathbf{M}^{(i)}\mathbf{v} = \mathbf{v}$.*

Proof. Equivalence may be verified directly. □

Lemma 4. *Suppose \mathbf{v} satisfies $\mathbf{M}^{(i)}\mathbf{v} = \mathbf{v}$. Then for all k we must have $|v_k| \leq |v_i|$.*

Proof. Suppose not. Then there exists k such that (i) $|v_k| > |v_i|$ and (ii) $|v_k|$ is maximal. Without loss of generality $\mathcal{P}(i, k) \neq \emptyset$ since otherwise $v_k = \sum_j M_{kj}^{(i)} v_j = \sum_j 0 v_j = 0$. Then $|v_m| = |v_k|$ for all $m \in H(\bullet, k)$, since $v_k = \sum_m M_{km}^{(i)} v_m = \sum_{m \in H(\bullet, k)} \frac{1}{|H(\bullet, k)|} v_m$ and $\sum_{m \in H(\bullet, k)} \frac{1}{|H(\bullet, k)|} = 1$. By induction on a path from i to k we arrive at $|v_i| = |v_k|$, contradicting (i). □

Lemma 5. *There exists a unique solution to $\mathbf{M}^{(i)}\mathbf{v} = \mathbf{v}$ with $v_i = 1$.*

Proof. Existence: The i th row of $\mathbf{M}^{(i)}$ contains only zeros, except for the i th entry which is equal to 1. Hence the characteristic polynomial contains a root (eigenvalue) $\lambda = 1$ with corresponding eigenvector $\tilde{\mathbf{v}} \neq \mathbf{0}$. From Lemma 4 we must have $v_i \neq 0$ (else $|\tilde{v}_j| \leq |\tilde{v}_i| = 0$ for all j , implying that $\tilde{\mathbf{v}} = \mathbf{0}$). Taking $\mathbf{v} = \tilde{\mathbf{v}}/v_{ii}$ proves existence.

Uniqueness: Suppose \mathbf{v} and \mathbf{v}' both satisfy the criteria. Then $\mathbf{v}'' = \mathbf{v} - \mathbf{v}'$ is also an eigenvector of $\mathbf{M}^{(i)}$ with unit eigenvalue. From Lemma 4 we have $|v_k''| \leq |v_i''| = 0$ for all k , implying that $\mathbf{v} = \mathbf{v}'$. \square

Proof of Theorem 4. Together Lemmas 1-5 prove that the effect matrix $\mathbf{e} = \{e_{ij}\}$ exists and is unique. Moreover, the i th row \mathbf{e}_i of \mathbf{e} may be computed as the unique eigenvector of $\mathbf{M}^{(i)}$ corresponding to unit eigenvalue.

For moderate dimensional matrices $\mathbf{M}^{(i)}$, eigenvectors may be calculated using any valid algorithm. One such algorithm, suitable for high dimensional problems, is the “power iteration” method. Specifically we compute the leading eigenvector \mathbf{v} of $\mathbf{M}^{(i)}$ as

$$\mathbf{e}_i = \lim_{n \rightarrow \infty} \mathbf{v}^{(n)}, \quad \mathbf{v}^{(n)} = \frac{\mathbf{M}^{(i)}\mathbf{v}^{(n-1)}}{\|\mathbf{M}^{(i)}\mathbf{v}^{(n-1)}\|}, \quad v_k^{(0)} = \delta_{ik}. \quad (32)$$

Efficient computation in this setting is reviewed in Berkhin (2005). \square

References

- Banerjee, A., Jost, J. (2009) Graph spectra as a systematic tool in computational biology. *Discrete Appl Math* **157**(10):2425-2431.
- Berkhin, P. (2005) A Survey on PageRank Computing, *Internet Mathematics* **2**(1):73-120.
- Breitkreutz *et al.* (2012) Molecular signaling network complexity is correlated with cancer patient survivability, *Proc Natl Acad Sci USA* **109**(23):9209-9212.
- Cantone *et al.* (2009) A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches, *Cell* **137**(1):172-181.
- Chuang *et al.* (2007) Network-based classification of breast cancer metastasis, *Mol Syst Biol* **3**:140.
- Dash, D. (2003) Caveats for Causal Reasoning with Equilibrium Models, PhD thesis, Intelligent Systems Program, University of Pittsburgh.
- Davis, J., Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*. ACM Press New York pp.233-240

- Drummond, C., Holte, R.C. (2004) What ROC curves can't do (and cost curves can). *Proc. 1st Workshop ROC Analysis in AI*, pp.19-26.
- Fawcett, T. (2006) An introduction to ROC analysis. *Pattern Recogn* **27**, 861-874.
- Feiglin *et al.* (2012) Static network structure can be used to model the phenotypic effects of perturbations in regulatory networks. *Bioinformatics* **28**(21):2811-2818.
- Feizi *et al.* (2013) Network deconvolution as a general method to distinguish direct dependencies in networks, *Nat Biotechnol* **31**:726-733.
- Goadrich, M., Oliphant, L., Shavlik, J. (2004) Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction. *Proceedings of the 14th International Conference on Inductive Logic Programming (ILP)*.
- Heiser *et al.* (2012) Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc Natl Acad Sci USA* **109**(8):2724-2729.
- Hill *et al.* (2012) Bayesian Inference of Signaling Network Topology in a Cancer Cell Line. *Bioinformatics* **28**(21):2804-2810.
- Iwasaki, Y., Simon, H.A. (1994) Causality and model abstraction. *Artif Intell* **67**(1), 143-194.
- Jeong *et al.* (2000) The large-scale organization of metabolic networks. *Nature* **407**(6804):651-654.
- Johannes *et al.* (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics* **26**(17):2136-2144.
- Jurman *et al.* (2011a) Biological network comparison via Ipsen-Mikhailov distance. arXiv:1109.0220.
- Jurman, G., Visintainer, R., Furlanello, C. (2011b) An introduction to spectral distances in networks. *Fr Art In* **226**:227-234.
- Lakhina *et al.* (2004) Structural analysis of network traffic flows. *Proceedings of International Conference on Measurement and Modeling of Computer Systems, ACM New York USA*, pp 61-72.
- Maathuis *et al.* (2010) Predicting causal effects in large-scale systems from observational data. *Nat Methods* **7**(4):247-248.
- Marbach *et al.* (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J Comp Biol* **16**(2):229-239.
- Marbach *et al.* (2010) Revealing strengths and weaknesses of methods for gene network inference. *Proc Natl Acad Sci USA* **107**(14):6286-6291.

- Marbach D, *et al.* (2012) Wisdom of crowds for robust gene network inference. *Nat Methods* **9**:796-804.
- Milenković *et al.* (2009) Optimized Null Models for Protein Structure Networks. *PLoS One* **4**(6):e5967.
- Morrison *et al.* (2005) GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics* **6**:233.
- Nelander *et al.* (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol* **4**(1):216.
- Oates, C., Mukherjee, S. (2012a) Network Inference and Biological Dynamics. *Ann Appl Stat* **6**(3):1209-1235.
- Page *et al.* (1999) The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford InfoLab*.
- Pearl, J. (2000) Causality: Models, Reasoning, and Inference. *Cambridge University Press, Cambridge UK*.
- Peters, J., Bühlmann, P. (2013) Structural Intervention Distance (SID) for Evaluating Causal Graphs. arXiv:1306.1043.
- Prill *et al.* (2010) Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges. *PLoS One* **5**(2):e9202.
- Prill *et al.* (2011) Crowdsourcing Network Inference: The DREAM Predictive Signaling Network Challenge. *Sci Signal* **4**(189):mr7.
- Scutari, M., Nagarajan, R. (2013) Identifying significant edges in graphical models of molecular networks. *Artif Intell Med* **57**(3):207-217.
- Simon, N., Tibshirani, R. (2012) Comment on “Detecting Novel Associations in Large Datasets” by Reshef *et al.*. *Technical Report, Stanford, CA, USA*.
- Städler, N., Mukherjee, S. (2013) Two-Sample Testing in High-Dimensional Models. arXiv:1210.4584.
- Tong *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science* **303**(5659):808-813.
- Wang *et al.* (2008) Motif-directed network component analysis for regulatory network inference. *BMC Bioinformatics* **9**(Suppl. 1):S21.
- Warshall, S. (1962) A theorem on Boolean matrices. *J ACM* **9**(1):11-12.

- Weile *et al.* (2012) Bayesian integration of networks without gold standards. *Bioinformatics* **28**(11):1495-1500.
- Winter, C. (2012) Google Goes Cancer: Improving Outcome Prediction for Cancer Patients by Network-Based Ranking of Marker Genes. *PLoS Comput Biol* **8**(5):e1002511.
- Yates, P.D., Mukhopadhyay, N.D. (2013) An inferential framework for biological network hypothesis tests. *BMC Bioinformatics* **14**(1):94.

B Supplement to “Quantifying the Multi-Scale Performance of Network Inference Algorithms”

B.1 AUC Convention

For MSS2 only, the analogue of the confusion matrix does not respect monotonicity of the entries TP, FP, TN, FN with respect to varying threshold parameter τ . As a consequence both the ROC and PR curves constructed by interpolation have the potential to exhibit self-intersection and hence there may exist estimates \hat{G} such that the area S under the associated ROC and PR curves (AUC) need not be bounded by $0 \leq S \leq 1$, invalidating the definition of a score function having image in $[0, 1]$. To construct a consistent performance measure we instead define AUC as follows: Let $(x_1, y_1) \rightarrow (x_2, y_2) \rightarrow \dots \rightarrow (x_m, y_m)$ denote the coordinates of points on the curve whose AUC is to be calculated. Reorder the sequence as $(x_{n_1}, y_{n_1}) \rightarrow (x_{n_2}, y_{n_2}) \rightarrow \dots \rightarrow (x_{n_m}, y_{n_m})$ such that the x-coordinates form an increasing sequence $x_{n_1} \leq x_{n_2} \leq \dots \leq x_{n_m}$. Then AUC is defined using the trapezium rule as $S = \sum_{i=2}^m \frac{y_{n_{i-1}} + y_{n_i}}{2} \times (x_{n_i} - x_{n_{i-1}})$. Note that whilst an ordering $\{x_{n_j}\}$ may in general be non-unique, the expression for S does not depend on which particular ordering is selected.

B.2 The “net assess” Package

The `net_assess` package for MATLAB R2013b is provided to compute all of the scores and associated p -values discussed in this paper. In the case of MSS2 we compute ROC and PR curves based on 100 uniformly spaced thresholds τ_i defined implicitly by $|E(\hat{G}^{\tau_i})| = (i/100) \times |E(\hat{G})|$. This package can be used as shown in the following example:

```
A2 = rand(100)>0.98; \% benchmark network
A1 = 0.5*rand(100) + 0.5*rand(100).*A2; \% estimate
score = 1; \% 1 = local, 2 = MSS1, 3 = MSS2
[AUROC,AUPR] = net_assess(score,A1,A2)
MC_its = 100; \% num. Monte-Carlo samples for p-val. computation
[AUROC,AUPR,p_AUROC,p_AUPR] = net_assess(score,A1,A2,MC_its)
```

Typical computational times for the package, based on the DREAM5 data analysis performed in this paper, are displayed in Supp. Fig. 8. Note that local scores require $\mathcal{O}(p^2)$ operations to evaluate, whereas MSS1 requires $\mathcal{O}(p^3)$ and MSS2 requires $\mathcal{O}(p^4)$.

B.3 Overall Score (Marbach *et al.*)

We briefly summarise the approach of Marbach *et al.* (2012) that assigns each network inference algorithm an “Overall Score”. For each network inference algorithm a combined AUROC score was calculated as the (negative logarithm of the) geometric mean over p -values corresponding to 3 individual AUROC scores (*in silico*, *E. Coli*, *S. Cerevisiae* respectively). Likewise a combined AUPR score was assigned to each network inference algorithm. For each algorithm an Overall Score was computed as the arithmetic mean of the combined AUROC

and combined AUPR scores. Note that Marbach *et al.* (2012) report the Overall Score over all 3 datasets, whereas the performance scores which we report in this paper were computed for each dataset individually. Note also that p -values used by Marbach *et al.* (2012) were computed using a common null distribution, rather than the estimator-specific nulls $\mathcal{M}_0(\hat{G})$ that are used in this paper.

B.4 Additional Results

Supp. Figs. 3, 4, 5 display the scores S obtained by each of the 36 dream methodologies on, respectively, the *in silico*, *E. Coli* and *S. Cerevisiae* datasets. Supp. Fig. 6 compares results for the DREAM5 Challenge data based on both local and multi-scale scores and ROC analysis. Supp. Fig. 7 displays both inferred benchmark topology for the *in silico* dataset using *Regression 2*.

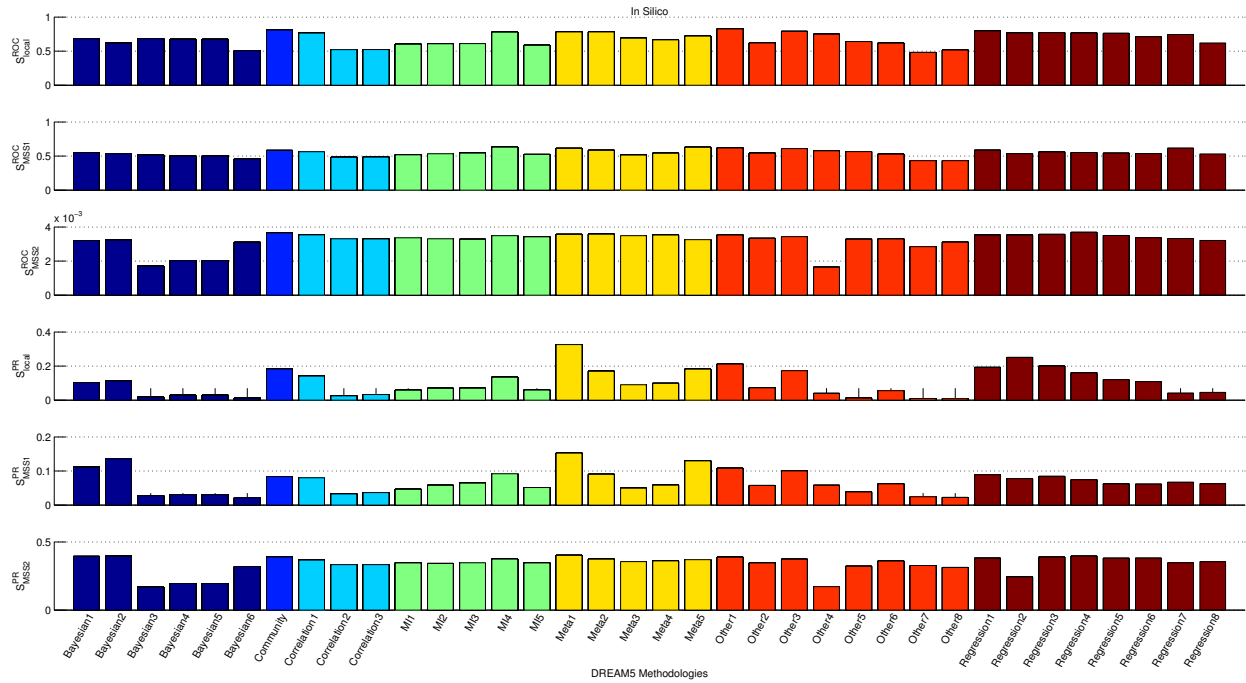


Figure 3: DREAM5 Challenge data; results on the *in silico* dataset. [The 36 DREAM5 methodologies were assessed using both local and multi-scale performance scores. Participants were grouped according to their statistical formulation as either *Bayesian*, *Correlation*, *Meta*, *Mutual Information* (MI), *Regression* or *Other*. *Community* represents a crowd-sourced network estimator proposed by Marbach *et al.* (2012). Performance scores include area under the receiver operating characteristic (AUROC) and precision recall (AUPR) curves, based on both local scores and the proposed multi-scale scores (MSS). The Overall Score of Marbach *et al.* (2012) combines both local AUROC and AUPR scores as described in Supp. Sec. B.3.]

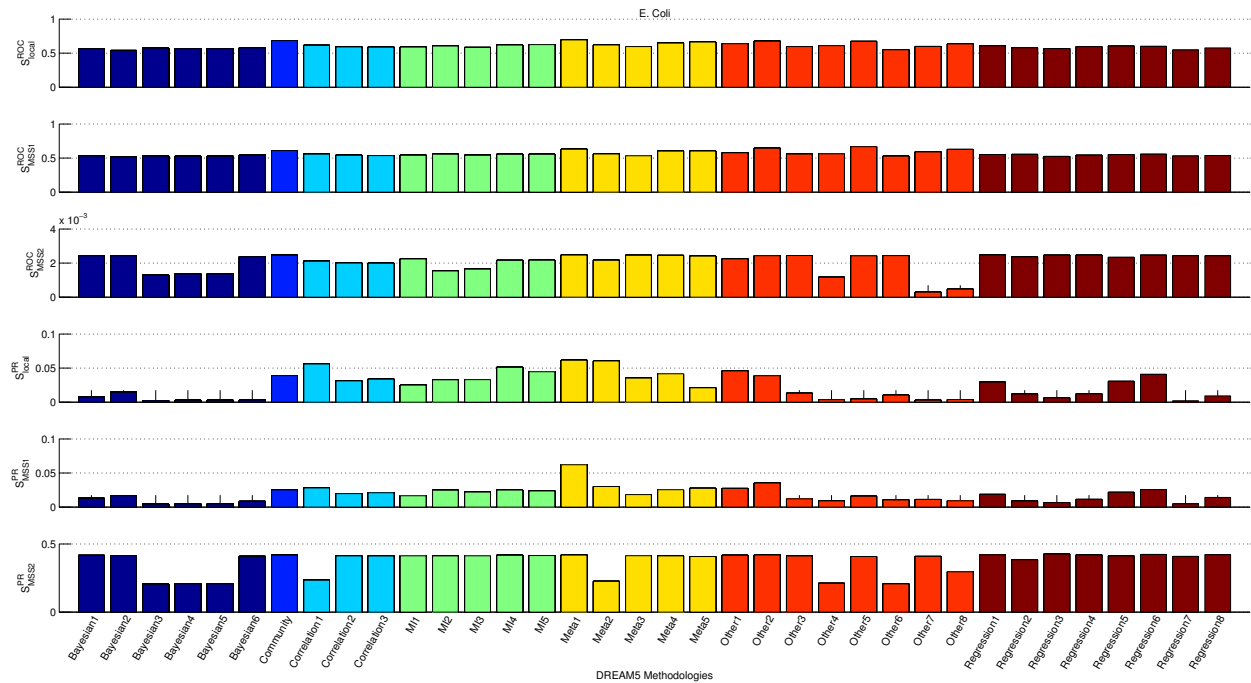


Figure 4: DREAM5 Challenge data; results on the *E. Coli* dataset. [The 36 DREAM5 methodologies were assessed using both local and multi-scale performance scores. Participants were grouped according to their statistical formulation as either *Bayesian*, *Correlation*, *Meta*, *Mutual Information* (MI), *Regression* or *Other*. *Community* represents a crowd-sourced network estimator proposed by Marbach *et al.* (2012). Performance scores include area under the receiver operating characteristic (AUROC) and precision recall (AUPR) curves, based on both local scores and the proposed multi-scale scores (MSS). The Overall Score of Marbach *et al.* (2012) combines both local AUROC and AUPR scores as described in Supp. Sec. B.3.]

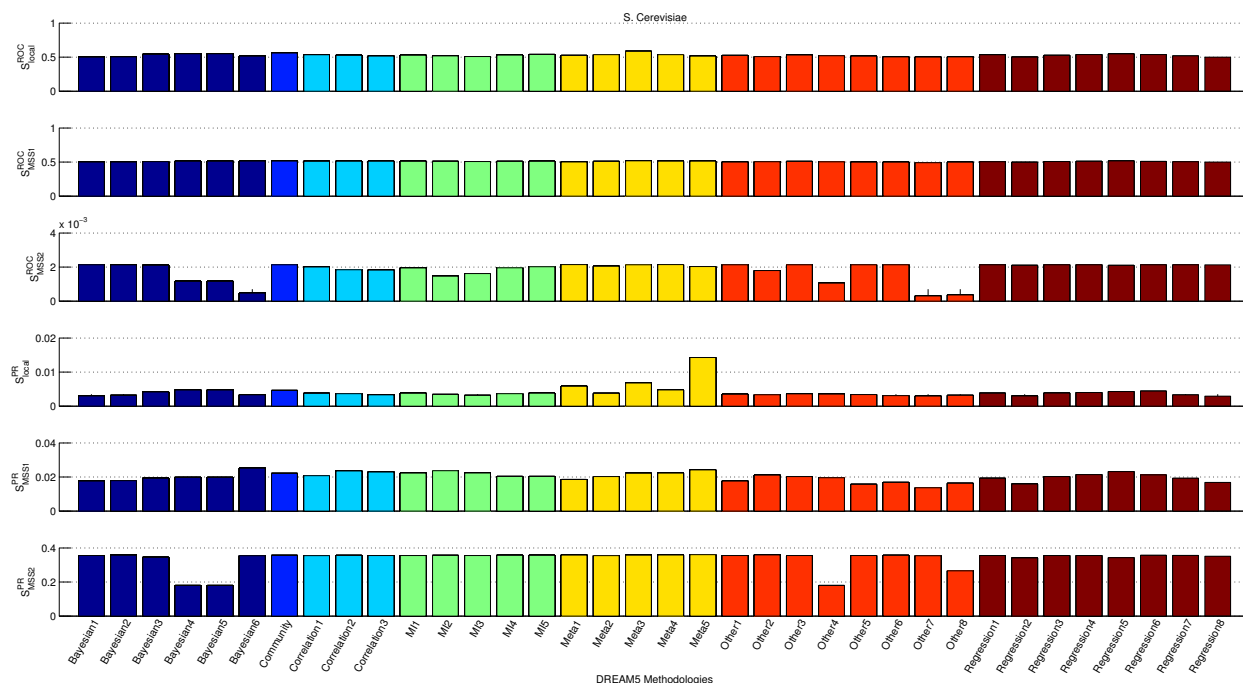


Figure 5: DREAM5 Challenge data; results on the *S. Cerevisiae* dataset. [The 36 DREAM5 methodologies were assessed using both local and multi-scale performance scores. Participants were grouped according to their statistical formulation as either *Bayesian*, *Correlation*, *Meta*, *Mutual Information* (MI), *Regression* or *Other*. *Community* represents a crowd-sourced network estimator proposed by Marbach *et al.* (2012). Performance scores include area under the receiver operating characteristic (AUROC) and precision recall (AUPR) curves, based on both local scores and the proposed multi-scale scores (MSS). The Overall Score of Marbach *et al.* (2012) combines both local AUROC and AUPR scores as described in Supp. Sec. B.3.]

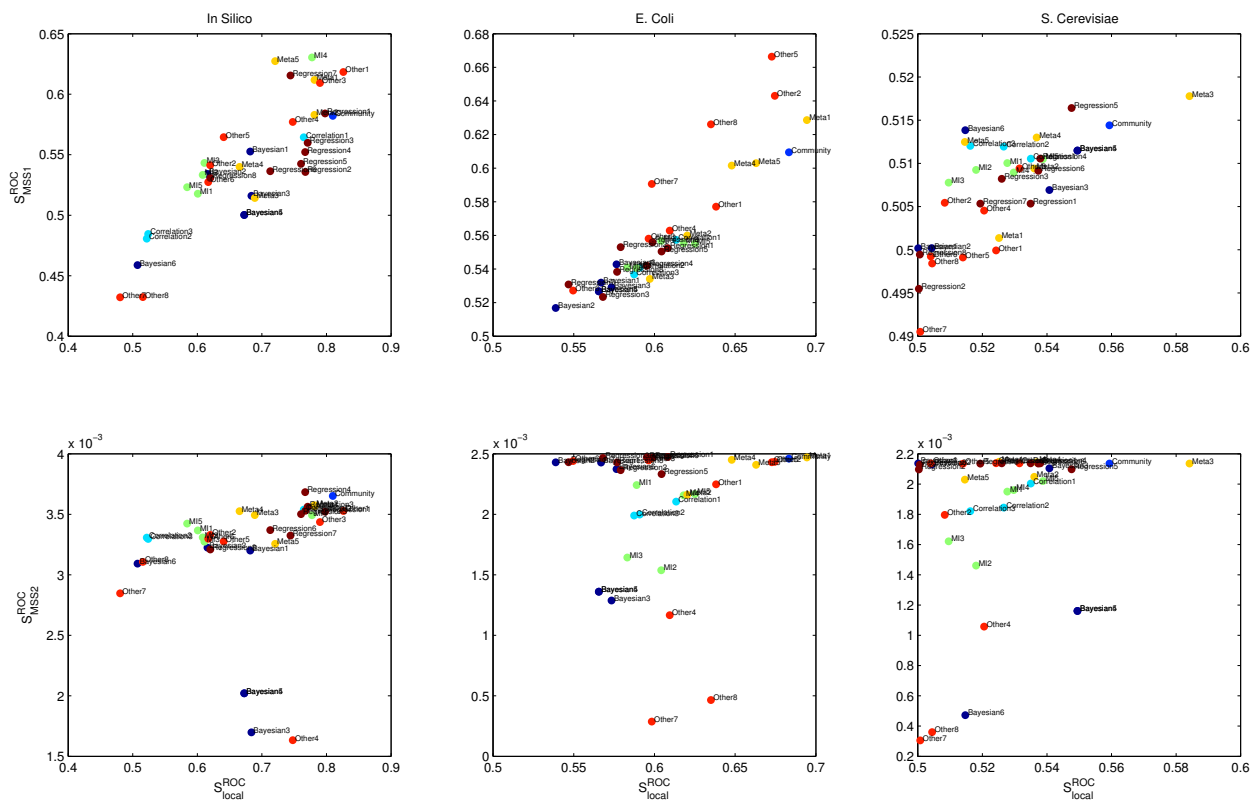
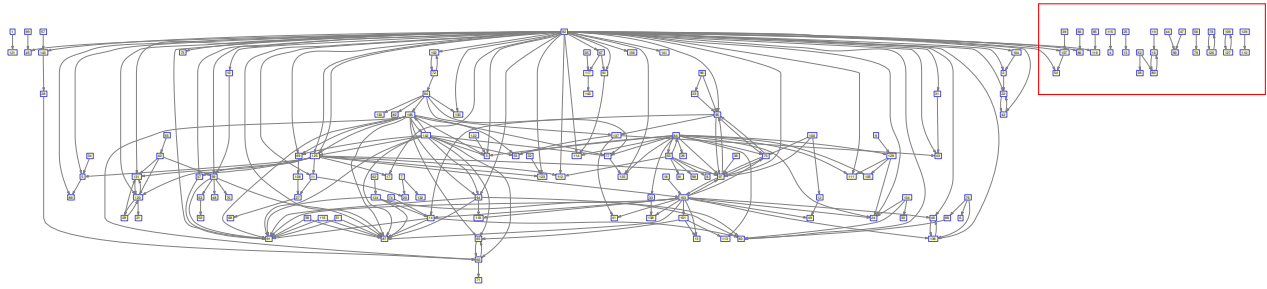
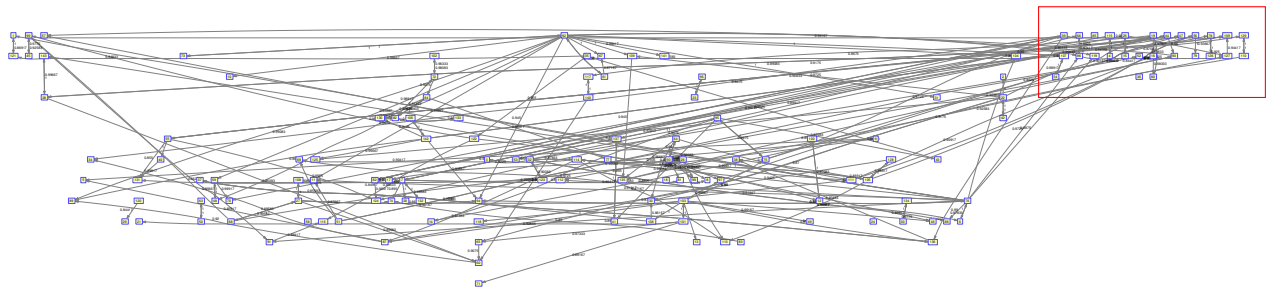


Figure 6: DREAM5 network inference challenge data. [The 36 DREAM5 methodologies were assessed using both local and multi-scale performance scores. Participants were grouped according to their statistical formulation as either *Bayesian*, *Correlation*, *Meta*, *Mutual Information* (MI), *Regression* or *Other*. *Community* represents a crowd-sourced network estimator proposed by Marbach *et al.* (2012). Left to right: *In silico* dataset, *E. Coli* dataset, *S. Cerevisiae* dataset. Here we show results based on area under the receiver operating characteristic curve.]



(a) *In silico* data-generating network



(b) Network topology inferred by *Regression2*

Figure 7: Global connectivity patterns in the DREAM5 *in silico* challenge; *Regression2* is unable to identify disconnected components (red box) in the data-generating network. [High definition available in electronic version.] (a) *In silico* data-generating network, based on known transcriptional regulatory networks of model organisms (Marbach *et al.*, 2009). (b) Network topology inferred by *Regression2*; an approach that combines steady-state and time-series data using a group LASSO, followed by bootstrapping. [Edges and associated weights are shown for the highest weights only, such that both networks (a) and (b) contain an equal number of edges.]

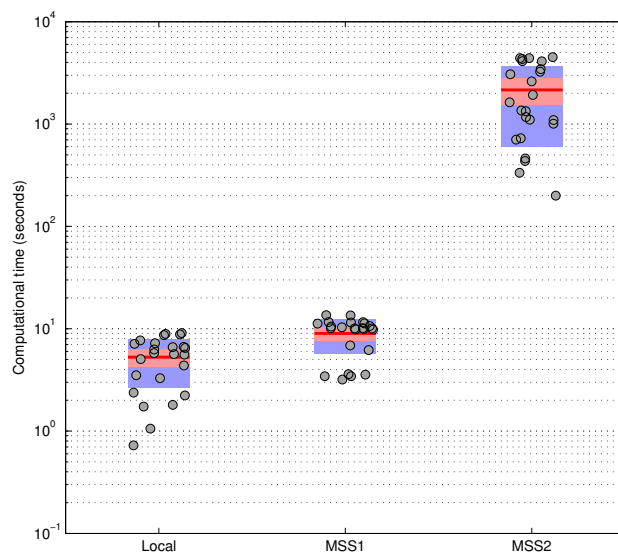


Figure 8: Computational times for the DREAM5 Challenge data. [Red; 95% confidence intervals for the mean. Blue; inter-quartile range. The full data are shown as points.]