Environmental Modelling and Software 133 (2020) 104796

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft

LINX: A topology based methodology to rank the importance of flow measurements in compartmental systems

Caner Kazanci [a,b,*], Malcolm R. Adams [a], Aladeen Al Basheer [a], Kelly J. Black [a], Nicholas Lindell [a], Bernard C. Patten [c], Stuart J. Whipple [c]

[a] Department of Mathematics, University of Georgia, Athens, GA 30602, USA
[b] College of Engineering, University of Georgia, Athens, GA 30602, USA
[c] Odum School of Ecology, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Keywords:
Energy–matter
Budget
Network model
Systems modeling
Input–output analysis
Link importance
Optimization

ABSTRACT

In ecological and other transactional energy–matter flow networks, accurate quantification of flows between compartments can be difficult and costly. For models at steady state or undergoing linear change, energy–matter conservation together with the steady-state condition can be exploited to estimate unknown flows from known ones. In compartmental network models, some flows are more important than others in terms of their connections to other flows, participation in cycles, geodesic distance to the environment (in the graph theoretical sense), and other topological features. In respect to estimating unknown flows, such importance differences also come into play. Pursuing this, we formulate a Link Importance iNdeX (LINX) that quantifies each flow's importance in a model. This index identifies and quantifies the redundancy imposed by network topology and mathematical conservation rules. We anticipate that it will find use in minimizing the cost and effort of data collection while also increasing model accuracy.

Software availability

Software name: LINX
Developer: Caner Kazanci, Kelly J. Black
System requirements: Linux, Mac OS, Windows
Program language: Matlab, C++
Availability: Matlab file exchange (https://www.mathworks.com/matlabcentral/fileexchange/72143-linx), GitHub (https://github.com/KellyBlack/LINX)
License: GPL-3.0

1. Introduction

In recent decades of ecological and other applied complex-systems modeling, numerous methods have been developed to quantify flow networks representing qualitative food webs. A typical approach to determining flows begins with a literature search for observational or experimental studies providing data for computation or estimation of model flows. References such as Jørgensen et al. (1991) that provide an annotated compendium of process rates with references may be quite useful. If the literature does not provide adequately reliable data, the next step often entails gathering empirical data at the model field site. In the food-web field, methods such as gut content analysis, stable isotope analysis (Pacella et al., 2013; Phillips and Gregg, 2003; Post, 2002), and fatty acid composition (Iverson et al., 2004) have been used to quantify dietary composition for use in network flow quantification.

Given the difficulty and high cost of gathering empirical data (Yodzis and Innes, 1992), methods based on inference and computation have been described. One such approach uses allometric principles derived from empirical observation of organism physiology and biochemical rate processes to calculate hypothetical flows in steady-state networks (Barnes et al., 2014). This method has deep roots in ecology and finds its greatest recent development in what has become known as metabolic theory in ecology (Brown et al., 2004; Sibly et al., 2012). Some network quantification procedures are based on community assembly rules, in which algorithms are used to build network structure and determine flows (Fath et al., 2007); these methods include a modified niche model (Halnes et al., 2007) and a structured food-web approach with network flows drawn from a probability distribution (Morris et al., 2005). Ulanowicz and Scharler (2008) describe two network quantification methods (joint apportionment and reverse mold-filling) that require minimal mathematical inference and avoid

[*] Corresponding author at: Department of Mathematics, University of Georgia, Athens, GA 30602, USA.
E-mail address: caner@uga.edu (C. Kazanci).

https://doi.org/10.1016/j.envsoft.2020.104796
Received 1 January 2020; Received in revised form 8 May 2020; Accepted 8 July 2020
Available online 22 July 2020

the complexities of optimization routines. Ecopath with Ecosim (Christensen and Walters, 2004) uses some of these methods to help construct and quantify a network of species interactions.

Finally, as flows (or ranges thereof) are determined, the model must be checked for consistency since conservation laws provide constraints which the measured flows may not satisfy. Specifically, the difference between the total input into and output from each compartment should be within expected range of values. For smaller systems this might be done by hand, using expert knowledge of the system, but mathematical optimization schemes, such as linear inverse modeling (LIM), introduced into ecology by Vézina and Platt (1988), have been developed considerably and applied in many ecological contexts (Breed et al., 2004; Saint-Béat et al., 2018; Marquis et al., 2007; Vézina and Pace, 1994; van Oevelen et al., 2010; Savenkoff et al., 2001).

Despite the existence of extensive literature focusing on computational methods that help quantify flows in network models, there is as yet no methodology to help guide and optimize the process of gathering empirical data. Approaches like LIM and balancing methods are helpful after empirical data collection, but are not designed to interact with the collection process itself. The same ecological principles and computational ideas that make such computational methods possible can be exploited to help optimize empirical data collection to lower the cost and effort, as well as to increase accuracy.

The methodology developed herein originated out of the observation that for steady-state network flow models, certain quantified flows may provide more or less information about other unquantified flows. The difference is determined solely by the location of a flow within the network structure. Given a list of compartments (nodes, vertices) and connections (links, flows), the method works by assigning importance values to each connection. The higher the importance value, the more the information gained by the quantification of that flow. Our method could be used in the context of any of the network flow quantification approaches described above where incomplete flow quantification leads to the question—which set of remaining unquantified flows should become focal for further empirical or analytical quantification?

Furthermore, it is noteworthy that certain flow magnitudes may provide information that results in a more stable calculation of the remaining, unquantified flows as well (see Section 6). That is to say, a small error in one measurement may propagate in different ways compared to the measurement of another flow in its place. One of the goals of our method is to determine which flows result in a lower potential variance of the calculations of the remaining flows in the presence of measurement errors. Applying our method in this way could improve the efficiency of model-making and aid in deploying research efforts in the construction of ecosystem flow networks.

Mathematical formulation of the *Link Importance iNdeX* (LINX) of this paper involves linear algebra. However, the general idea motivating it can be described verbally, as is done in Section 2 using a three compartment food-chain model. A slightly more complicated model is needed to show the relevance and usefulness of a LINX, and understand how it works. We use a simple model with three compartments and five flows to do this in Section 4, after covering necessary assumptions and notations in Section 3. A general LINX formulation is derived in Section 5, and refined and finalized in Section 6. Section 7 includes a demonstration using the oyster reef ecosystem model (Dame and Patten, 1981), with emphasis on practical considerations such as data availability. Computational feasibility is discussed in Section 8, followed by the conclusion (Section 9).

2. Flow importance index: The idea

In a network model at steady-state, not every flow needs to be quantified empirically, because the steady-state assumption introduces constraints. In Fig. 1, for example, determining any one of the four flows is sufficient to quantify the remaining three flows because the steady-state condition forces all the flows to be equal. Their importance values are also equal because the amount of information each provides about the others is identical.

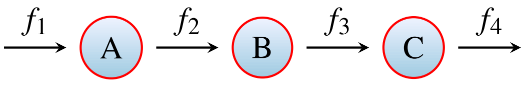

Fig. 1. A three compartment food chain.

Because each flow determines all the others, a chain model is too simple to fully motivate a generalized importance-index concept. The example in Section 4 will demonstrate, using a more complicated network, how some flow values can be more useful than others in determining unknown values. In that section we will introduce the link importance index, LINX, that will quantify the usefulness of each of the flows. Flows with higher LINX values will give more restrictions on the whole system than those with lower indices. We preferred “link importance index” to “flow importance index” because the mathematical formulation of LINX is independent of the flow values. It is based purely on network topology, the structure and organization of connections, or links among compartments.

LINX does not only help identify the most helpful flows to build a complete model, but also provides information about model accuracy. In theory, determining all four flow values in the model shown in Fig. 1 would be totally unnecessary and redundant, as the values must all be equal. In broader practice, however, depending on the amount of error involved in determining an individual flow, averaging the four measured flow values in this example should increase accuracy because it effectively quadruples the sample size. In general, a higher LINX value will imply that accurate measurement of that flow will contribute significantly more to the overall model accuracy than a flow with lower LINX value. In situations where restrictions on cost and effort limit the utilization of highly accurate measurement methods for the entire flow data, the guidance provided by the flow importance index can be quite useful.

The source of the extra information provided by the measurement of certain flows is due to the steady-state assumption that the total amount of input received by a compartment equals its total output per unit time. In fact, LINX is still applicable to certain models not at steady-state. Appropriate mathematical representation of a flow network is necessary to formulate the effect of the steady-state assumption, and how it can be relaxed in LINX usage, as next discussed.

3. Notation and the steady-state assumption

It is common to represent quantified flows in steady-state multi-compartment models by matrices. The representation of the flow orientation differs in literature. For instance, Patten (1978) employs a columns(j)-to-rows(i) flow orientation, f_{ji}, to represent the flow from compartment i to j. Ulanowicz (1986) represents the same flow using a rows(i)-to-columns(j) flow orientation, T_{ij}. Since the environment is not represented as a compartment in standard network analyses (e.g. Patten and Ulanowicz), environmental inputs into or outputs from compartments are not included in the flow matrix, but are represented separately by vectors. The environmental input into and output from compartment i are denoted by z_i and y_i, respectively, by Patten (1978). Ulanowicz (1986) denotes environmental inputs as D_i, and distinguishes two categories of environmental outputs, usable (export, E_i) and unusable (respiration, R_i).

The mathematical representation of the steady-state assumption is significantly simpler when all flows, including the environmental inputs and outputs, are enumerated and represented in vector form instead. For that reason, Vézina and Platt (1988) forego the use of matrices entirely, and instead use a vector r, where r_i represents the ith flow. We adopt this convention for present purposes, however we use a different letter, f_i, in keeping with the notation of Patten.

The information imposed by the steady-state assumption on a food chain ($f_i = f_j$ for all i, j) is quite simple, which is not necessarily the case for a general ecosystem model. In order to accurately identify the information imposed by the steady-state assumption on a general ecosystem model, we need a formal way to represent how flows and compartments are connected to each other for a given ecosystem model. In graph theory, the incidence matrix is used for this purpose. Since open ecosystem models are not really graphs,[1] we use a generalized version of the incidence matrix, called the stoichiometric matrix, which is the same as the incidence matrix for closed systems with no environmental inputs or outputs.

The *stoichiometric matrix* S (Eq. (1)), along with flow vector f, provides the simplest mathematical representation of the conservation laws (Eq. (3)). For a model with n compartments and k flows (including the environmental inputs and outputs), the stoichiometric matrix (Resendis-Antonio, 2013) has n rows and k columns, and is defined as follows:

$$S_{ij} = \begin{cases} 1 & \text{if flow } f_j \text{ is to compartment } i, \\ -1 & \text{if flow } f_j \text{ is from compartment } i, \\ 0 & \text{neither.} \end{cases} \tag{1}$$

For example, the stoichiometric matrix of the three compartment food chain in Fig. 1 is

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix}, \tag{2}$$

where flows and compartments are ordered as $\{f_1, f_2, f_3, f_4\}$ and $\{A, B, C\}$, respectively. The vector f contains flow magnitudes, ordered in the same way as the columns of the stoichiometric matrix. For example, the second column of S contains -1 on its first row and $+1$ on its second row, representing the flow from the first compartment (A) to the second (B). In other words, when a unit flow occurs from A to B, there will be a unit decrease in A, a unit increase in B, and no change in C. Thus, the second column of S, $[-1, 1, 0]$ corresponds to the second flow f_2. The fact that a model is at steady-state can be represented in terms of the matrix S and the vector f as

$$Sf = 0. \tag{3}$$

For example, for the three-compartment food chain in Fig. 1, this simple equation yields the correct steady-state conditions as follows:

$$Sf = 0 \iff \begin{matrix} f_1 - f_2 = 0 \\ f_2 - f_3 = 0 \\ f_3 - f_4 = 0 \end{matrix} \iff f_1 = f_2 = f_3 = f_4.$$

In general, for an n-compartment model, the steady-state assumption provides n linear equations involving flow values that can be exploited to determine unknown flows without needing to quantify all of them.

While the steady-state assumption is essential for the method to work, it is not absolutely needed. The same approach still works if the change in compartment storage values can be assumed to occur at a constant rate. For example, if a certain compartment gains (or loses) approximately a certain amount of storage per unit time over a time period, then all the methods presented in this paper remain applicable during that time period. In general, if the rate of change of the storage value of compartment i is given as c_i, then S and f satisfy the following equation instead of Eq. (3).

$$Sf = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$$

This does not apply to systems where storage values fluctuate significantly, which can be characterized by the high values of the magnitudes of the second derivative of the storage values, over the observed time period. We will continue to use the phrase "steady-state assumption" throughout this manuscript for simplicity and clarity. Nevertheless, the entire content of this manuscript remains valid for models where changes in storage values can be assumed to occur at constant rates.

4. Flow importance index: A simple example

The three compartment food chain model (Fig. 1) is too simple an example to demonstrate how a link importance index is to be formulated in general. Material covered in Section 3 enables us to demonstrate how the flow importance measure is computed for the simple three compartment model shown in Fig. 2. Unlike the previous example, no single flow value is enough to determine all five flows in this model. We need to find the minimum number of flows that need to be quantified in order to compute all flows in this system. For this, consider the steady-state equations for each compartment:

$$\begin{aligned} A &: f_1 = f_2 + f_3 \\ B &: f_2 + f_4 = f_5 \\ C &: f_3 = f_4 \end{aligned} \tag{4}$$

This linear system has five variables (degrees of freedom) and three equations (restrictions), leaving two degrees of freedom. This means if we fix two of the flows, the remaining three should be uniquely determined. That is, quantifying only two flows will usually suffice to determine all five flows of this model. To do this explicitly, we rewrite the steady-state equations (4) using the stoichiometric matrix, as shown in Eq. (3).

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{5}$$

To demonstrate, we can randomly choose two of the five flows, and compute the remaining three based on the two we chose. As an example, we choose f_4 and f_5, and compute f_1, f_2 and f_3 in terms of f_4 and f_5. To do this, we need a system of equations that expresses f_1, f_2 and f_3 in terms of f_4 and f_5. This follows from Eq. (5):

$$\begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = - \begin{bmatrix} 0 & 0 \\ 1 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} f_4 \\ f_5 \end{bmatrix}$$

Here, the 3×3 matrix on the left is made from the first three columns of the stoichiometric matrix (4) that correspond to the unknown flows f_1, f_2 and f_3, and the 3×2 matrix on the right is formed from the remaining columns of S that correspond to f_4 and f_5. Then, assuming the 3×3 matrix on the left is invertible, we get the equations we need to compute f_1, f_2 and f_3, demonstrating that quantifying f_4 and f_5 is sufficient to determine all flow-rates:

$$\begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = - \begin{bmatrix} 1 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 1 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} f_5 \\ -f_4 + f_5 \\ f_4 \end{bmatrix} \tag{6}$$

This exercise informs us about the necessary and sufficient condition that the matrix formed by the columns of S that correspond to the

[1] For open ecosystem models, each environmental input and output is connected to a single compartment, as the environment is not included within the boundaries of the model. For a graph however, each flow (edge) needs to be attached to exactly two compartments (vertices), by definition. *Hypergraphs* do not have this restriction. Also called *set systems* (Hell and Nešetřil, 1970), they are significantly more general mathematical constructs than graphs, and therefore not the most suitable abstraction for the ecosystem models we focus on.

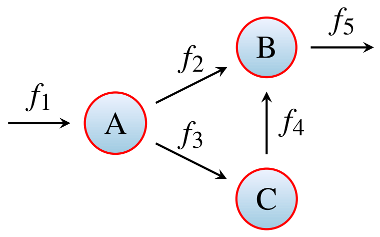

Fig. 2. A three compartment ecosystem model with five flows.

unknown flows should be invertible. If this matrix were not invertible, it would be impossible to derive the equations we need, implying that it would be impossible to uniquely compute f_1, f_2 and f_3 using f_4 and f_5. To demonstrate such a case, if we try to find f_2, f_3 and f_4 in terms of f_1 and f_5, we get

$$\begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} f_2 \\ f_3 \\ f_4 \end{bmatrix} = - \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_5 \end{bmatrix}. \tag{7}$$

In this, the 3 × 3 matrix on the left is not invertible, meaning that knowing f_1 and f_5 is not sufficient to determine f_2, f_3 and f_4. The intuitive reason for this is that measuring both f_1 and f_5 is redundant, as the steady-state assumption implies that $f_1 = f_5$. Similarly, $f_3 = f_4$. Any other choice of two flows however, can be used to determine all five flows in the model.

The goal of LINX is to rank the amount of information on other flow values gained by quantifying a specific flow. The extra information is afforded by the steady-state assumption, and the amount of this information depends on the location of the compartment within the network and the entire network topology. As an example, for the simple food-chain model of Fig. 1, measuring any one of the four flows does provide the values of the three remaining flows as well, effectively quadrupling the information provided by a single measurement. This unusually high gain is due to the steady-state condition combined with the fact that each compartment has exactly one input and one output, forcing all flows to be equal to each other. LINX indicates that for this model, quantifying any one of the flows provides the same amount of information.

Assessing the contribution of a single flow for the three compartment model in Fig. 2 is more complicated, as at minimum two flows are needed to compute all five flows. Only if we quantify a flow in question and another flow will it then be possible to compute all other flows, given that the matrix formed by the columns of the stoichiometric matrix corresponding to the remaining three flows is invertible (6). If that matrix is not invertible (7), then those two flows alone will not be enough to determine all flows in the system. Based on this observation, we define LINX as the proportion of the cases that makes it possible to compute all five flows, when a flow in question is paired with another flow. This statement is generalized (8), formulated (10), and improved (13) in the following sections.

As an example, we compute the flow importance index of f_3. There are four different sets of two flows that contain f_3 and one other flow: $\{f_1, f_3\}, \{f_2, f_3\}, \{f_3, f_4\}, \{f_3, f_5\}$. Considering the invertibility of the resulting matrices formed by each of these sets (Eqs. (6) and (7)), we see that all five flows in the model can be computed by measuring the flows in any of these sets, except for $\{f_3, f_4\}$. In other words, out of the four flow sets that have the potential to compute all five flows, only three of them can be used to do so. Therefore, we define the flow importance index of f_3 as 3/4. We can compute the importance indices of all flows by first making a list of all sets containing two different flows:

$$\{f_1, f_2\}, \{f_1, f_3\}, \{f_1, f_4\}, \cancel{\{f_1, f_5\}}, \{f_2, f_3\}, \{f_2, f_4\}, \{f_2, f_5\}, \cancel{\{f_3, f_4\}}, \{f_3, f_5\}, \{f_4, f_5\}.$$

The two cross-outs contain redundant flows. One can use the two flows in any of the eight remaining sets to compute all five flows.

We define the *link importance index (LINX)*, $m(f_i)$, as the fraction of sets with two elements including f_i that can be used to quantify all flows. Previously, we observed that out of the four sets containing f_3, only three can be used to quantify all flows. The same statement is valid for f_1, f_4 and f_5. However, f_2 is different in that all four groups containing f_2 can be used to quantify all flows. The flow importance indices for Fig. 2 model are:

$$m(f_1) = m(f_3) = m(f_4) = m(f_5) = \frac{3}{4} = 0.75, \; m(f_2) = \frac{4}{4} = 1.$$

The value of $m(f_2)$ is higher than the others because once we know the value of f_2, quantifying any one of the remaining four flows is sufficient to derive all flows. This is not true for any of the other flows, implying that quantifying f_2 before others, or more accurately than others, is advantageous, which is the goal of LINX.

This section has illustrated how LINX is computed for a simple model. In the next section we derive a preliminary general mathematical formulation for LINX, and in Section 6 we improve this formulation to give the full definition of LINX.

5. LINX: Preliminary general formulation

To construct a preliminary general formulation for LINX, we need first to find out the minimum number of flows needed to determine all flows in multi-compartment models in general. For an n-compartment model, which always contains $k > n$ flows, including environmental inputs and outputs, at least $k-n$ flows must be quantified to determine all flows. This is because the steady-state condition forms a linear system with k variables and n equations, leaving $k - n$ degrees of freedom (see Appendix A for proof.) However, as in the above example, not every set of $k - n$ flows will work. This observation leads to our preliminary formulation that is given by

$$m(f_i) = \frac{\text{Number of flow sets of size } k-n \text{ including } f_i\text{, sufficient to determine all flows}}{\text{Number of flow sets of size } k-n \text{ including } f_i}. \tag{8}$$

We formulate the denominator first. The number of sets containing $k - n$ distinct flows is the number of subsets of the set of all flows $\mathcal{F} = \{f_1, \ldots, f_k\}$ containing $k-n$ elements. This, in turn, is the number of combinations of k flows taken $k-n$ at a time:

$$\binom{k}{k-n} = \frac{k!}{(k-n)!\,n!}.$$

Since f_i has to be one of the elements, we need to count the combinations of the remaining $k - 1$ flows taken $k - n - 1$ at a time, which is

$$\frac{(k-1)!}{(k-n-1)!\,n!} = \binom{k-1}{n}.$$

Next we formulate the numerator, which is the number of the sets counted in the denominator that can be used to determine all the flows. For instance, to compute $m(f_3)$ for the model depicted in Fig. 2 of the previous section, first identify all sets containing two flows, one of which is f_3: $\{f_1, f_3\}$, $\{f_2, f_3\}$, $\{f_3, f_4\}$, $\{f_3, f_5\}$. The number of such sets is $\binom{5-1}{3} = 4$, which is the denominator of the LINX definition in Eq. (8). The numerator of the preliminary LINX definition only counts the sets that allow the computation of all flows. To find out if this is the case for a given set, we need to form the linear equations that express the undetermined flows in terms of the quantified flows, as we did in Eqs. (6) and (7). For example, let us derive the system of linear equations for the first set $\mathcal{A} = \{f_1, f_3\}$ the steady-state condition

$Sf = 0$:

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

This linear equation should express the unknown flows $\{f_2, f_4, f_5\}$ in terms of the known flows $\{f_1, f_3\}$ and test if a unique solution exists. In order to do so, we need to divide the matrix S and vector f into two parts accordingly.

$$\begin{bmatrix} 1 & -1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_3 \end{bmatrix} + \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} f_2 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{9}$$

In general, let $\mathcal{F} = \{f_1, f_2, \dots, f_k\}$ represent the set of all flows. For a subset $\mathcal{A} \subset \mathcal{F}$, we define $f(\mathcal{A})$ as a vector containing only the flows in $\mathcal{A}$. Similarly, we define $S(\mathcal{A})$ as a matrix formed by the columns of S that correspond to the flows in $\mathcal{A}$. Then a general version of Eq. (9) is

$$S(\mathcal{A})f(\mathcal{A}) + S(\mathcal{F}\backslash\mathcal{A})f(\mathcal{F}\backslash\mathcal{A}) = 0$$

where the set $\mathcal{F}\backslash\mathcal{A}$ contains all flows that are not in $\mathcal{A}$ (i.e., the complement of $\mathcal{A}$). For our specific example $\mathcal{F}\backslash\mathcal{A} = \{f_2, f_4, f_5\}$. Then the general form of the linear equation that represents the unknown flows $(\mathcal{F}\backslash\mathcal{A})$ in terms of the known flows $(\mathcal{A})$ is

$$f(\mathcal{F}\backslash\mathcal{A}) = -[S(\mathcal{F}\backslash\mathcal{A})]^{-1} S(\mathcal{A}) f(\mathcal{A}).$$

For our specific example, this equation is

$$\begin{bmatrix} f_2 \\ f_4 \\ f_5 \end{bmatrix} = - \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} 1 & -1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_3 \end{bmatrix}$$

which can be simplified as

$$\begin{bmatrix} f_2 \\ f_4 \\ f_5 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} f_1 - f_3 \\ 0 \\ f_3 \end{bmatrix}.$$

The flows $\{f_1, f_3\}$ can be used to determine all five flows if this equation is solvable, which only possible if the shown 3 × 3 matrix is invertible. In general, the flows in set $\mathcal{A}$ can be used to uniquely determine all flows $(\mathcal{F})$ if and only if the matrix $S(\mathcal{F}\backslash\mathcal{A})$ is invertible, in other words, $\det(S(\mathcal{F}\backslash\mathcal{A})) \neq 0$. This observation is all we need to construct a general formula for $m(f_i)$, which is the proportion of all flow sets of size $k - n$ containing f_i, which can be used to compute all flows

$$m(f_i) = \frac{1}{\binom{k-1}{n}} \sum_{\substack{\mathcal{A}\subset\mathcal{F} \\ f_i\in\mathcal{A} \\ |\mathcal{A}|=k-n}} \operatorname{sgn}|\det(S(\mathcal{F}\backslash\mathcal{A}))|\,. \tag{10}$$

In the expression under the sum, the symbol $|\mathcal{A}|$ represents the number of elements (cardinality) of $\mathcal{A}$. Thus, the sum is over the subsets $\mathcal{A}$ of $\mathcal{F}$ that have $k - n$ elements, and contain the ith flow, f_i. The number of such sets is $\binom{k-1}{n}$, which is the denominator. For the numerator, we should count only the sets $\mathcal{A}$ with $\det(S(\mathcal{F}\backslash\mathcal{A})) \neq 0$. We do that by taking the absolute value and sign function of the determinant, so that it equals one if this condition is satisfied, and zero otherwise,

$$\operatorname{sgn}|\det(S(\mathcal{F}\backslash\mathcal{A}))| = \begin{cases} 1, & \text{if } S(\mathcal{F}\backslash\mathcal{A}) \text{ is invertible,} \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

By this preliminary definition, LINX always takes values between zero and one. If the LINX value of a flow is one, then it would certainly be a good idea to quantify that flow first, whereas flows with low LINX values should be the last ones to be quantified. For most well-connected ecosystem models, LINX will usually be less than one-half, unlike the values we observed for the simple models we used for demonstration.

6. LINX: Improved formulation

Computation of unknown flows based on quantified flows requires the solution of a linear system of equations, $Ux = v$, where v is a vector based on quantified flows, U is a matrix based on the model's network structure, and x represents the unknown flows to be determined. For the three-compartment model shown in Fig. 2, one of these linear equations is

$$\underbrace{\begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & -1 \\ 0 & -1 & 0 \end{bmatrix}}_{U=S(\mathcal{F}\backslash\mathcal{A})} \underbrace{\begin{bmatrix} f_2 \\ f_4 \\ f_5 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} f_3 - f_1 \\ 0 \\ -f_3 \end{bmatrix}}_{v}. \tag{12}$$

One issue with the solution of such equations is the propagation of error from the quantified flows (f_1, f_3) to the computed solution (f_2, f_4, f_5). In ideal conditions, one might naively expect that 10% error in f_1, f_3 would imply a 10% error on f_2, f_4, f_5. In reality, the error in f_2, f_4, f_5 can be as high as, but cannot exceed 37.3%. This "error amplification factor" of 3.73 of a linear equation system is determined by the *condition number* of its matrix U, which is computed as the ratio of its maximum singular value to its minimum singular value (Beezer, 2015).

An $n \times n$ linear system will be under-determined and will not have a unique solution if the corresponding matrix U is not invertible. The condition number of a non-invertible matrix is ∞, meaning that any small error can be infinitely magnified. Matrices that are very close to being non-invertible have high condition numbers, meaning that the solutions they provide have high potential to magnify the errors in quantified empirical data. This fact is relevant to LINX, as it is based on the ability of a given flow, when combined with others, to determine all the flows in a model. Determining the entire flow data in a model requires the solution of a linear system, like Eq. (12). The preliminary definition of LINX (10) is entirely based on whether or not this linear system is solvable (11), but not how accurate the resulting numerical approximation is for the whole system. In other words, it does not factor in how the existing error in quantified flows will propagate to the computed flows. If, for a given set of flows, the condition number is high, the computed flows will not be nearly as reliable, decreasing the value of the additional information provided by quantifying those focal flows. Accurate assessment of this additional information is precisely the goal of LINX; therefore, the additional information provided by the condition number should be factored into its formulation. This is relatively straightforward to accomplish:

$$\text{Original:}\quad m(f_i) = \frac{1}{\binom{k-1}{n}} \sum_{\substack{\mathcal{A}\subset\mathcal{F} \\ f_i\in\mathcal{A} \\ |\mathcal{A}|=k-n}} \operatorname{sgn}|\det(S(\mathcal{F}\backslash\mathcal{A}))|\,.$$

$$\begin{aligned} &\text{Revised:}\quad m(f_i) = \frac{1}{M} \sum_{\substack{\mathcal{A}\subset\mathcal{F} \\ f_i\in\mathcal{A} \\ |\mathcal{A}|=k-n}} \frac{1}{\kappa(S(\mathcal{F}\backslash\mathcal{A}))}, \\ &M = \max_j \left\{ \sum_{\substack{\mathcal{A}\subset\mathcal{F} \\ f_j\in\mathcal{A} \\ |\mathcal{A}|=k-n}} \frac{1}{\kappa(S(\mathcal{F}\backslash\mathcal{A}))} \right\}. \end{aligned} \tag{13}$$

The improved formulation is different in two ways. First, the expression inside the sum is replaced with the multiplicative inverse of the condition number, κ, of the relevant matrix. The expression $\operatorname{sgn}|\det(S(\mathcal{F}\backslash\mathcal{A}))|$ takes the value 1 or 0 depending on the matrix $S(\mathcal{F}\backslash\mathcal{A})$ being invertible or not, respectively. If $S(\mathcal{F}\backslash\mathcal{A})$ is not invertible, then $1/\kappa(S(\mathcal{F}\backslash\mathcal{A})) = 1/\infty = 0$ as well, matching the preliminary definition. Unlike the preliminary definition, however, $1/\kappa(S(\mathcal{F}\backslash\mathcal{A}))$ often takes values significantly lower than one if $S(\mathcal{F}\backslash\mathcal{A})$ is invertible. For example, Table 1 shows these values for the three-compartment model in Fig. 2. For the three sets with non-zero determinants, the inverse of the condition numbers are 0.25, 0.28 and 0.38. This poses a

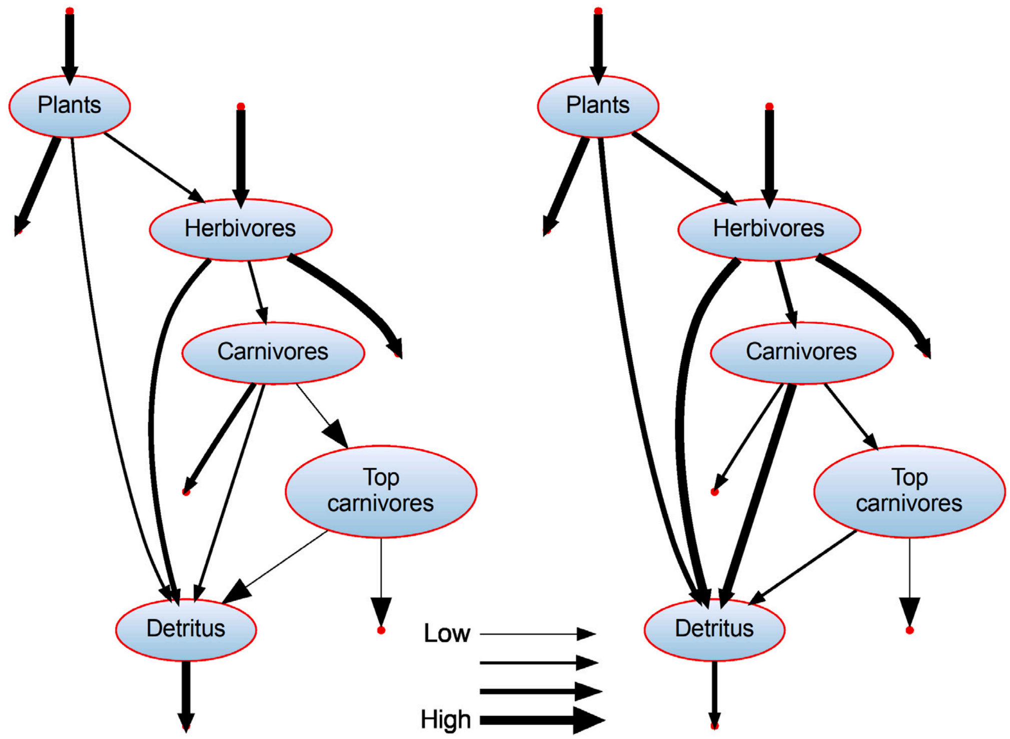

Fig. 3. LINX values of the Silver Springs energy flow model (Odum, 1957; Kemp and Boynton, 2004) is shown using both the preliminary and improved formulation for comparison. For better visualization, LINX values are indicated by using different line thicknesses. The thicker a flow line, the higher its LINX value. The red dots indicate the environment (Kazanci, 2007; Schramski et al., 2011).

Table 1
Detailed computation of $m(f_3)$ using the preliminary and improved formulations (13) for Fig. 2 model. The first column lists all four sets (of size 2) containing f_3. The second column uses the preliminary formulation, and assigns a value of 1 if the two flows (in the first column) can be used to determine all five flows of the model, and assigns a value of 0 otherwise. The third column uses the improved formulation and assigns values between 0 and 1, depending on how accurately all five flows can be computed. Sums of these values are smaller for the improved formulation, and therefore different scaling factors are used. The single and double underlined values indicate the preliminary and improved importance indices of f_3, respectively.

$\mathcal{A}$	$\text{sgn}\lvert\det(\mathcal{S}(\mathcal{F}\setminus\mathcal{A}))\rvert$	$1/\kappa(\mathcal{S}(\mathcal{F}\setminus\mathcal{A}))$
$\{f_1, f_3\}$	1	0.28
$\{f_2, f_3\}$	1	0.38
$\{f_3, f_4\}$	0	0
$\{f_3, f_5\}$	1	0.25
Sum_{f_3}	3	0.91
$\text{Sum}_{f_3}/\binom{5-1}{3}$	$\underline{0.75}$	0.23
$\text{Sum}_{f_3}/\max_j\{\text{Sum}_{f_j}\}$	0.75	$\underline{\underline{0.71}}$

problem, as just replacing the expression inside the sum, the improved importance index would take the value $(0.28 + 0.38 + 0 + 0.25)/4 = 0.23$ whereas the preliminary formulation gives $(1 + 1 + 0 + 1)/4 = 0.75$. To compensate for the characteristic low values of the improved formulation, we do not divide the sum by the number of sets $\binom{k-1}{n}$, but by the maximum such sum. This way, the preliminary and improved importance indices achieve comparable values, as shown on Table 2.[2] While the proportion of the values are roughly preserved, the additional information provided by the condition numbers placed the flows f_3 and f_4 ahead of f_1 and f_5, whereas all four flows had the same importance value according to the older definition.

Table 2
Comparison of LINX values for the Fig. 2 model computed using the preliminary (10) and improved (13) formulations.

	$m(f_1)$	$m(f_2)$	$m(f_3)$	$m(f_4)$	$m(f_5)$
Preliminary formulation	0.75	1	0.75	0.75	0.75
Improved formulation	0.61	1	0.71	0.71	0.61

Fig. 3 shows a more realistic comparison of the improved versus preliminary LINX formulations using a classical ecological energy-flow model (Odum, 1957; Kemp and Boynton, 2004). This model has $n = 5$ compartments and $k = 14$ flows, so there exist $\binom{14-1}{5} = 1287$ flow groups containing any specific flow. In Fig. 3, the thickness of a flow line indicates its importance. The clear difference between the two diagrams demonstrates the significant effect of condition numbers on the importance of flows, which was not observed in the simple three-compartment model (Table 1.) Even though the Silver Springs model is small in size, its slightly more complicated network topology compared to the three-compartment model causes a difference significant enough to change the order of importance of flows when the improved formulation is used. For example, according to the preliminary definition, the environmental output of the "Carnivores" compartment (Carnivores → Environment) is more important than the flow from "Carnivores" to "Detritus" (Carnivores → Detritus.) On the other hand, it is exactly the opposite when the improved formulation is used. Even though the improved formulation is slightly more resource intensive to compute, from here on we will employ it by default, as it provides a more accurate measure of link importance. The following two sections discuss several issues that may arise in using LINX for model building, refining, or tuning. We address some of these with a view to effective application of the methodology.

[2] Actually, for large systems, the preliminary LINX will also generally produce importance indices significantly less than one. Thus, in general, to compare the preliminary index with the improved index it would be best to scale both of them so that the largest importance value is 1.

7. Practical considerations: Data availability

Quantifying a flow not only informs about its value, but values of other flows as well, because of the steady-state assumption. LINX simply computes the amount of this additional information to determine the importance of each link. This information cannot be accurately computed for flows about which prior information exists from previous observations, experiments, or literature. Such empirical information will render LINX information partially useless. Also, known flow values will provide information on other flows as well, which may significantly alter LINX values.

It is not just known flow values that affect LINX values. If a model contains flows that are difficult to measure accurately,[3] then any additional information that can help quantify those flows will be commensurately valuable. Such difficult-to-measure flows can have a significant effect on the LINX values as well. This section explains how LINX is computed in these two situations using the three-compartment model of Fig. 2, and illustrates the ideas using an intertidal oyster reef ecosystem model (Fig. 4.)

We consider three scenarios that can arise about flow f_1 in determining LINX values in Fig. 2 model:

1. f_1 is already known via pre-existing data, or because it is the focus of the research question, and so have a high importance by default;
2. It is not feasible to determine f_1 empirically, and no data for it exists; and
3. f_1 is known, and it is not feasible to determine f_2 accurately (combination of the first two cases).

For simplicity, in this example only, we will use the preliminary LINX formulation. It is possible to compute the LINX values using the preliminary formulation simply by counting, whereas the condition numbers of several matrices need to be computed in the case of the improved formulation. In practice, the improved formulation is preferred due to its accuracy, therefore it will be used for the intertidal oyster-reef model analysis presented at the end of this section (Fig. 4).

Recall that we defined $m(f_i)$ for the three-compartment model in Section 4 as the fraction of sets with two flows ($k - n = 5 - 3 = 2$) including f_i, that can be used to quantify all flows. Listing all flow sets with two elements

$\{f_1, f_2\}, \{f_1, f_3\}, \{f_1, f_4\},$ ~~$\{f_1, f_5\}$~~$, \{f_2, f_3\}, \{f_2, f_4\}, \{f_2, f_5\},$ ~~$\{f_3, f_4\}$~~$,$
$\{f_3, f_5\}, \{f_4, f_5\}$

and striking out the two unusable ones (Section 4), we computed all five LINX values to be $m(f_2) = 1$ and $m(f_i) = 0.75$, $i = 1, 3, 4, 5$. For the three scenarios listed above, the importance index can be constructed using this same basic idea. Table 3 lists the relevant sets for each of the three different scenarios, as well as the importance indices computed using them. For the first scenario, f_1 is already known, so it should participate in all flow sets of size two, which reduces the number of such sets from ten to only four. It is exactly the opposite in the case of the second scenario, where f_1 cannot participate in any of the sets, since it is assumed that it is not feasible to quantify f_1 using direct methods such as field experiments or literature review. Scenario 3 is a combination of the previous cases, so f_1 should be in all sets, and f_2 cannot be in any of the sets. The computation of importance indices of all flows is done according to the same definition above, using the preliminary formulation, and results are shown on Table 3.

Table 3
LINX values are computed for the three-compartment model shown in Fig. 2 for the following three different scenarios: *(i)* f_1 is known via pre-existing data, *(ii)* it is not feasible to determine f_1 accurately, and *(iii)* f_1 is known, and it is not feasible to determine f_2 accurately. The preliminary formulation of LINX is used due to its simplicity, for demonstration purposes.

	Scenario 1	Scenario 2	Scenario 3
Relevant flow sets	$\{f_1, f_2\}$, $\{f_1, f_3\}$, $\{f_1, f_4\}$, ~~$\{f_1, f_5\}$~~.	$\{f_2, f_3\}$, $\{f_2, f_4\}$, $\{f_2, f_5\}$, ~~$\{f_3, f_4\}$~~, $\{f_3, f_5\}$, $\{f_4, f_5\}$.	$\{f_1, f_3\}$, $\{f_1, f_4\}$, ~~$\{f_1, f_5\}$~~.
$m(f_1)$	Known	Infeasible	Known
$m(f_2)$	1/1	3/3	Infeasible
$m(f_3)$	1/1	2/3	1/1
$m(f_4)$	1/1	2/3	1/1
$m(f_5)$	0/1	3/3	0/1

Comparing the LINX values for each of the three scenarios, we observe some similarities as well as significant differences. For example, the LINX value f_3, originally $m(f_3) = 0.75$, changes to 1, 0.67 and 1 for each of the three scenarios, respectively. The LINX value of f_5 on the other hand, originally $m(f_5) = 0.75$, changes to 1, 0 and 1 for each of the three scenarios, respectively. Drastic changes in $m(f_5)$ under different scenarios show that prior information and data availability can affect importance of flows significantly, and hence, should be taken into account. This observation is not specific to this small and simple model, and the preliminary formulation of LINX. Fig. 4 shows four different prior-information and data-availability scenarios for an intertidal oyster reef ecosystem using the improved formulation of LINX.

8. Practical considerations: Computational feasibility

In this section, we describe how to compute the LINX values using Matlab, and then discuss issues of performance and feasibility. A Matlab code that automatically computes the LINX values given a model's stoichiometric matrix is available at GitHub (Kazanci and Black, 2020) and Matlab Central/File Exchange (Kazanci, 2020). This code is compatible with the freely available GNU Octave software (Eaton et al., 2014) as well as Matlab. For instance, one can compute the LINX values for the three-compartment model of Fig. 2 using only its stoichiometric matrix (Eq. (5)) as follows:

```
>> S = [ 1 -1 -1 0 0; 0 1 0 -1 -1; 0 0 1 1 0 ];
>> m = LINX( S );
Number  Flow   LINX (-1:known -2:infeasible):
  1     * -> 1  0.6057
  2     1 -> 2  1.0000
  3     1 -> 3  0.7130
  4     2 -> 3  0.7130
  5     2 -> *  0.6057
```

The results match the ones provided in Table 2. Use of the stoichiometric matrix for ecosystem modeling is not as common as the adjacency matrix. A second Matlab code named `A2S.m` is available with `LINX.m` that automatically converts a given adjacency matrix that includes environmental inputs and outputs to the stoichiometric matrix. Therefore the same results shown above can be obtained by executing the command `LINX(A2S(A))`, where A is the adjacency matrix of the model.

It is also possible to use the same code in case of prior information, or if some flows are difficult to obtain. For instance, the LINX values corresponding to the three scenarios provided on Table 3 can be computed as follows, by simply listing the flows that are known or are infeasible to compute:

[3] An ecological example of a difficult-to-measure flow would be a feeding flow for a rare predator that is known to exist in the model, but is so rare that obtaining a consumption rate for it is nearly impossible, even considering the range of empirical methods available, such as direct observation or gut-content analysis, or indirect methods such as metabolic estimates or stable isotope analysis.

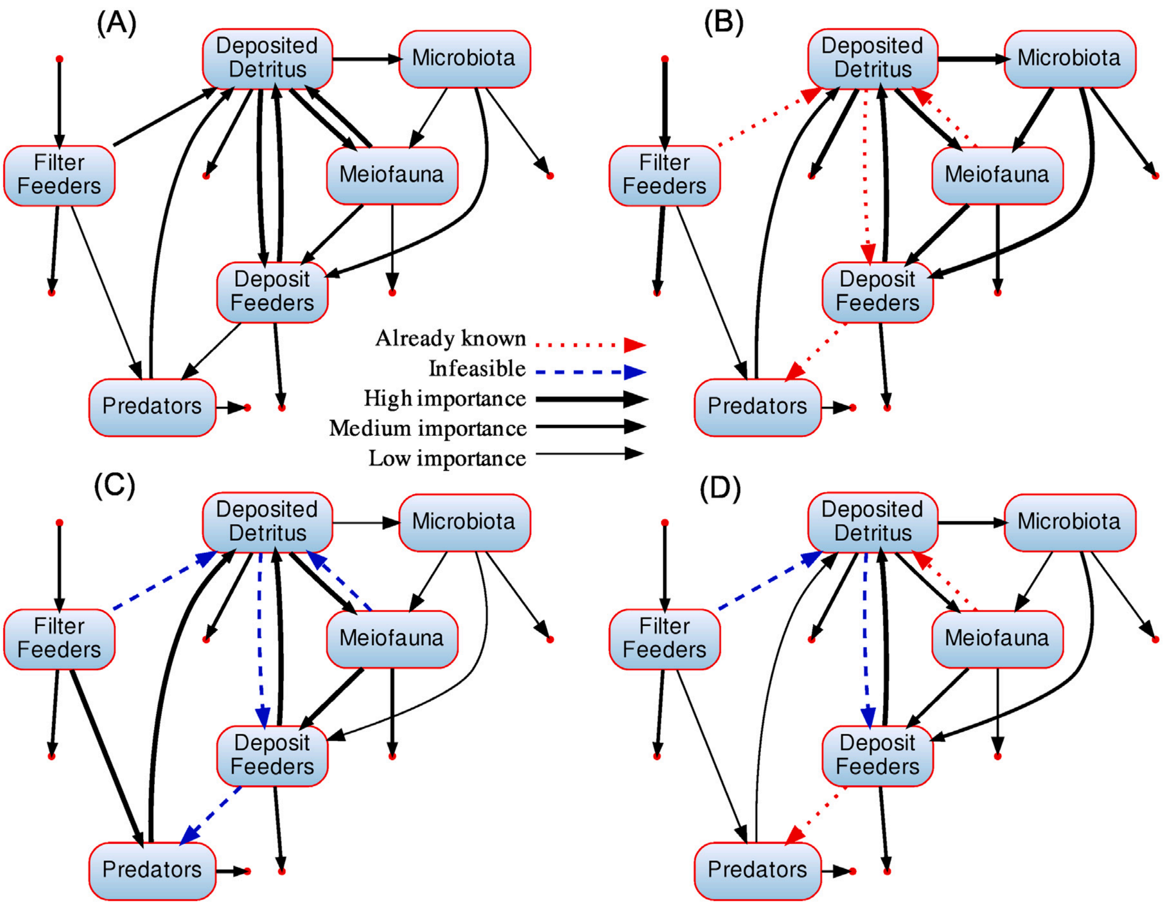

Fig. 4. LINX values are indicated for each flow of an oyster reef energy-flow model (Dame and Patten, 1981). Thicker arrows indicate higher LINX values. Figure (A) involves no prior information. Figure (B) assumes prior knowledge of four flow rates, indicated by red dotted lines. Figure (C) assumes the same four flows cannot be accurately determined, indicated by blue dashed lines. Figure (D) assumes two flows with prior information (red dotted) and two (blue dashed) that cannot be accurately determined. The LINX values are ordered differently under each scenario.

```
>> S = [ 1 -1 -1 0 0; 0 1 0 -1 -1; 0 0 1 1 0 ];
>> m = LINX( S, [1], [] );
Number  Flow    LINX (-1:known -2:infeasible):
  1      * -> 1   -1.0000
  2      1 -> 2   0.9217
  3      1 -> 3   1.0000
  4      2 -> 3   0.9217
  5      2 -> *   0.0000
>>m = LINX( S, [], [1] );
Number  Flow    LINX (-1:known -2:infeasible):
  1      * -> 1   -2.0000
  2      1 -> 2   1.0000
  3      1 -> 3   0.6222
  4      2 -> 3   0.6429
  5      2 -> *   0.7537
>>m = LINX( S, [1], [2] );
Number  Flow    LINX (-1:known -2:infeasible):
  1      * -> 1   -1.0000
  2      1 -> 2   -2.0000
  3      1 -> 3   1.0000
  4      2 -> 3   0.9217
  5      2 -> *   0.0000
```

These results are slightly different than provided on Table 3 because in the latter case the preliminary formulation of LINX was used for its simplicity. The results presented here use the improved formulation.

Figs. 3 and 4 are automatically generated using Graphviz. Graphviz uses a text file with “dot” extension to create images in multiple formats. A Matlab file named LINXdiagram.m that automatically generates this text file, which then can be used to create a network diagram, is provided at GitHub (Kazanci and Black, 2020) and Matlab Central/File Exchange (Kazanci, 2020).

One limitation of this code is its slowness for large models. This is because all possible combinations of flows must be examined, which is a computationally expensive process. It is possible to make use of the row-reduced echelon form (RREF) of the stoichiometric matrix to limit the number of combinations to be tested. This more complicated algorithm scales better for larger models, and further details are provided in Appendix B. A faster C++ code based on this algorithm is provided along with the Matlab code at GitHub (Kazanci and Black, 2020).

In the case of extremely large models, even the faster C++ code might not be feasibly utilized. It is not possible to provide a specific threshold however, as network topology and prior flow information (Section 7) have a significant effect on speed. In general, models with high connectances (Dunne et al., 2002) are expected to run slower, and models with prior information will definitely run faster. For example the number of flow combinations that needs to be tested for the oyster reef ecosystem model shown in Fig. 4 is 27132. Under the three different scenarios shown on the same figure, the number of combinations decrease to 5005, 105 and 1365 for cases B, C and D, respectively. In other words, knowledge that four of the nineteen flows are unlikely to be determined empirically, makes the algorithm about 270 times faster, in theory. In practice, the oyster reef model is small enough that all cases run almost instantly on a regular personal computer. This happens to be case for models of similar size. The difference can be drastic for large or complex models, however.

Model size and complexity are decidedly limiting factors in the applicability of the present LINX methodology to realistically complex systems studies, such as those undertaken in large continuing studies like the US National Science Foundation's Long Term Ecological Research (LTER) and National Ecological Observatory Network (NEON) programs. Even at such ambitious research scales, our LINX methodology could provide guidance for targeting empirical research to quantify flows in incomplete model flow networks as they are created, quantified, and revised over the life of projects.

In most cases large complex ecosystem models are the result of long-term team-based research projects, and empirical research would provide prior information on some of the model flows. Since prior information often provides for feasible application of the LINX algorithm, many of these models could use this methodology. Large models typically contain model sectors, which are groups of compartments that share many characteristics, such as microautotrophs, microheterotrophs, or macroheterotrophs. Since teams of researchers are often responsible for a certain model sector, such sectors might have some flows quantified and could be analyzed with the LINX algorithm as sub-system matrices, as well as the full model being analyzed at various stages in the model-building process.

9. Conclusion

Computational methods serve as an essential part of compartmental network-flow modeling. Their utilization, however, usually occurs after parameters have been determined through empirical data collection, missing data methods and literature search. The computational method of this paper is applied early in the parametrization phase of modeling, after some flow data, but not all, have been acquired. Exploiting conservation laws and network topology, LINX enables modelers to make informed decisions that guide the data collection, helping to build a more accurate model while minimizing resources necessary to determine flow values.

The usefulness of LINX is not just limited to before data collection. As Fig. 4 demonstrates, quantifying four flows produces system-wide large changes in the LINX values of the subsequent partially-quantified flow network. In other words, the quantification of a few focal flows can dramatically change the LINX values of other flows. The process described in Section 7 can be utilized repeatedly during the model-building process, and can identify which remaining flows are going to be redundant, eliminating further costly and potentially unnecessary empirical measurements. Furthermore, modelers may utilize redundant flows to increase the accuracy of other flows that may contain relatively greater errors. In other words, they can use LINX to identify which specific flows will help increase the accuracy of the others.

Finally, the significant effect of prior knowledge about specific flows on the LINX values of other flows reflects on the organization of large, complex, self-assembling systems like ecosystems. From the way our present algorithm works it is possible to infer that the importance structure of flows in ecosystems is very fluid and changing indeed in response to other network flows extant or not extant at any given time or place.

CRediT authorship contribution statement

Caner Kazanci: Conceived the ideas, Developed and refined the formulation, Software, Writing - original draft. **Malcolm R. Adams:** Conceived the ideas, Developed and refined the formulation, Writing - original draft. **Aladeen Al Basheer:** Conceived the ideas, Developed and refined the formulation, Writing - original draft. **Kelly J. Black:** Conceived the ideas, Developed and refined the formulation, Software, Writing - original draft. **Nicholas Lindell:** Contribution to the proof of the theorem, Writing - original draft. **Bernard C. Patten:** Writing - original draft. **Stuart J. Whipple:** Conceived the ideas, Developed and refined the formulation, Writing - original draft.

Acknowledgment

All authors participated in the preparation of the publication, contributed critically to the drafts and gave final approval for publication.

Appendix A. Minimum number of flows required to determine all flows

Theorem 1. *The rank of the $n \times k$ stoichiometric matrix S of any open and connected ecosystem model equals n, the number of its compartments. In other words, its rows are linearly independent, and so are n of its columns.*

Proof. First we note that $n \leq k$ (i.e., there are at least as many flows as there are compartments) since the system is connected and open. Also $\text{rank}(S) \leq \min(n,k) \leq n$ by definition. Assume for contradiction that $\text{rank}(S) < n$; then there is a nontrivial linear combination of the rows $\mathbf{r}_m$ such that

$$\sum_{m=1}^{n} c_m \mathbf{r}_m = \mathbf{0}, \quad \text{but } c_m \neq 0 \text{ for some } m. \tag{A.1}$$

Note that each of the k flows is attached to either one or two compartments. By construction, then, each column of S has either one or two nonzero entries. Since the system is open, there is at least one column with only one nonzero entry. Let the ith column be one of these. It corresponds to an environmental flow into or out of some compartment s, corresponding to the sth row , $\mathbf{r}_s$, of the matrix. This flow is not connected to any other compartment. We use induction on the length of the shortest path from compartment b to s to arrive at a contradiction as follows:

The ith entry of every row but $\mathbf{r}_s$ is zero, necessitating $c_s = 0$. Inductively, assume any compartment a with shortest path length L from s has $c_a = 0$. Let compartment b have distance $L+1$ to compartment s; certainly, b must be connected via flow j to some a which satisfies the inductive hypothesis. Only $\boldsymbol{r}_a$ and $\boldsymbol{r}_b$ have nonzero entries in column j, so that $c_a = 0 \implies c_b = 0$ also.

The system is connected, whence every coefficient is zero by the induction above, contradicting $c_m \neq 0$ for some m. Thus, $\text{rank}(S) = n$. □

Remark. This theorem shows that there exist collections of n flows such that the corresponding columns of S are linearly independent (and thus the $n \times n$ matrix formed by these columns is invertible). However, as evident from examples in the paper, it is often not the case that *every* collection of n columns yields an invertible matrix.

Appendix B. A faster algorithm

The minimum number of flows required to determine all flows is established within the main body of the primary paper. One immediate task is to determine an algorithm to determine which combinations of the flows determine the feasible sets. In other words, which set of n flows can be used to calculate all of the remaining flows. We will call these sets *acceptable.* Here we define two methods for searching the acceptable sets of flows. The first is a brute force method, and the second makes use of the row reduced echelon form (RREF) of the stoichiometric matrix to limit the number of combinations that are tested. Codes for both approaches are publicly available (Kazanci and Black, 2020).

The first method is to simply examine all possible n element sets of flows. An advantage of this method is it makes use of standard libraries readily available to calculate the combinations to test. Another advantage is that it is complete and robust. Every combination is tested and none will be missed. The shortcoming is that the algorithm does not scale well to larger systems, and a system with a larger number of nodes may require a long time to search through all possible combinations.

The approach to examine all possible combinations is a computationally expensive process, so another approach is explored that is designed to reduce the operation count. The primary disadvantage of the alternate method is that its implementation is more complicated. The basic idea is that the RREF of the stoichiometric matrix can be used to more narrowly define which columns of the stoichiometric matrix are linearly dependent. To better describe the approach, we first describe the problem given in Section 3 of the original paper. As a reminder, the linear system for the example can be expressed in the form

$$S=\begin{bmatrix}1 & -1 & 0 & 0\\ 0 & 1 & -1 & 0\\ 0 & 0 & 1 & -1\end{bmatrix}\begin{bmatrix}f_1\\ f_2\\ f_3\\ f_4\end{bmatrix}=0. \tag{B.1}$$

We note that Eq. (B.1) represents a system in the form

$$Sf=0. \tag{B.2}$$

In general, the matrix S has n rows and k columns, the vector f has k rows, and the zero vector on the right hand side has n rows. An alternate representation for the system is to express S in terms of its columns and the vector f as

$$\begin{bmatrix}s_1 & s_2 & \cdots & s_k\end{bmatrix}\begin{bmatrix}f_1\\ f_2\\ \vdots\\ f_k\end{bmatrix}=0. \tag{B.3}$$

Here the column vector s_i represents the entries in the stoichiometric matrix corresponding to whether or not there is an inflow or outflow for node i for a given flow. The system can be expanded to express it in an equivalent form,

$$f_1s_1+f_2s_2\cdots+f_ks_k=0. \tag{B.4}$$

The goal is to test all sets of n linearly independent vectors from s_i. The linear system given in Eq. (B.2) can be put in row reduced echelon form. This will transform the system where the s_i can be more easily compared as a way to rule out which columns might be expressed in terms of another column.

We first examine the example system given in Eq. (B.1). The system with the RREF of the matrix is given by

$$\begin{bmatrix}1 & 0 & 0 & 0 & -1\\ 0 & 1 & 0 & 1 & -1\\ 0 & 0 & 1 & -1 & 0\end{bmatrix}f=0$$

The task is to determine all sets that contain three column vectors that represent a linearly independent set. The first thing to notice is that the first row in the RREF matrix implies that s_1 and s_5 must be equal. Unless there is a nonzero entry in s_5 after the first row then there is no need to check a set of vectors that have both s_1 and s_5. The search for the first vector in the candidate sets can be conducted by using the first row and searching all sets where the first vector is s_1 or the first vector is s_5. Given these two initial vectors for a candidate set, the second row can then be searched. Because there are zeros in the second row for columns 1 and 3, a set of candidates for the second vector in the acceptable sets include the vectors s_2, s_4, or s_5. Finally, based on the third row, the only non-zero entries are in columns three and four, so candidates for the acceptable sets include those sets where the third vector is either vectors s_3 or s_4.

The general algorithm makes direct use of the row reduced echelon form of Eq. (B.2). The recursive algorithm proceeds by first testing sets where the first vector is one of columns that has a non-zero coefficient in the first row of the RREF form of the stoichiometric matrix. Each column in the second row that has a non-zero coefficient is then included for the second element of the possible sets. The algorithm then proceeds through each of the following rows until all possible sets of the columns have been tested.

Compared to the exhaustive method described above, this algorithm reduces the number of columns to test. The expense, though, is the use of a recursive algorithm that adds additional overhead. Another downside of the algorithm is that it requires additional book keeping and checking to insure that previously used combinations are not reused.

The C++ code requires that the stoichiometric matrix be defined in a regular text file. The first lines in the text file consist of the stoichiometric matrix. Each line is a row in the matrix, and the numbers are separated by spaces. Two optional lines can be appended below the stoichiometric matrix. One line can be used to define which flows are known in advance (known). The other line can be used to specify which flows cannot be measured (unknown). An example of a file is given below, where it is assumed that the flows for nodes 1, 3, and 6 are known while the flows for nodes 2, 5, and 11 cannot be determined.

```
1 -1 -1 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0
0 1 0 -1 -1 -1 0 0 1 0 1 0 1 0 -1 0 0 0 0
0 0 0 1 0 0 -1 -1 0 0 0 0 0 0 0 -1 0 0 0
0 0 0 0 1 0 1 0 -1 -1 0 0 0 0 0 0 -1 0 0
0 0 0 0 0 1 0 1 0 1 -1 -1 0 0 0 0 0 -1 0
0 0 1 0 0 0 0 0 0 0 0 1 -1 0 0 0 0 0 -1
known: 3 6 1
unknowable: 2 5 11
```

When the C++ program is run the name of the program (optimalFlow) is used to start the program. The name of the file to read is provided on the command line after the program name. For example, if the file shown above is named "oyster.txt" then the command to run the program is "opimtimalFlow oyster.txt" The program will read the file and determine approximations to the impact for each node. An example of the output for the text file above is given below:

```
Unknowable: 3-6-12
Known: 2-4-7
Number Flow   Impact
 1   *->1   0.82301
 2   1->2   (known)
 3   1->6      NA (unknowable)
 4   2->3   (known)
 5   2->4   0.74658
 6   2->5      NA (unknowable)
 7   3->4   (known)
 8   3->5   0.55752
 9   4->2   0.71440
10    4->5   0.80692
11    5->2   1.00000
12    5->6      NA (unknowable)
13    6->2   1.00000
14    1->*   0.82301
15    2->*   0.84232
16    3->*   0.34674
17    4->*   0.61062
18    5->*   0.87450
19    6->*   0.85438
Press <RETURN> to close this window...
```

References

Barnes, A.D., Jochum, M., Mumme, S., Haneda, N.F., Farajallah, A., Widarto, T.H., Brose, U., 2014. Consequences of tropical land use for multitrophic biodiversity and ecosystem functioning. Nature Commun. 5, 5351.

Beezer, R.A., 2015. A First Course in Linear Algebra. Congruent Press, Gig Harbor, Washington, USA, Open source.

Breed, G.A., Jackson, G.A., Richardson, T.L., 2004. Sedimentation, carbon export and food web structure in the mississippi river plume described by inverse analysis. Mar. Ecol. Prog. Ser. 278, 35–51.

Brown, J.H., Gillooly, J.F., Allen, A.P., Savage, V.M., West, G.B., 2004. Toward a metabolic theory of ecology. Ecology 85 (7), 1771–1789.

Christensen, V., Walters, C.J., 2004. Ecopath with Ecosim: methods, capabilities and limitations. Ecol. Model. 172 (2–4), 109–139.

Dame, R.F., Patten, B.C., 1981. Analysis of energy flows in an intertidal oyster reef. Mar. Ecol. Prog. Ser. 5 (2), 115–124.

Dunne, J.A., Williams, R.J., Martinez, N.D., 2002. Food-web structure and network theory: the role of connectance and size. Proc. Natl. Acad. Sci. 99 (20), 12917–12922.

Eaton, J.W., Bateman, D., Hauberg, S., Wehbring, R., 2014. GNU Octave Version 3.8.1 Manual: a High-Level Interactive Language for Numerical Computations. CreateSpace Independent Publishing Platform, ISBN: 1441413006.

Fath, B.D., Scharler, U.M., Ulanowicz, R.E., Hannon, B., 2007. Ecological network analysis: network construction. Ecol. Model. 208 (1), 49–55. http://dx.doi.org/10.1016/j.ecolmodel.2007.04.029.

Halnes, G., Fath, B.D., Liljenström, H., 2007. The modified niche model: Including detritus in simple structural food web models. Ecol. Model. 208 (1), 9–16. http://dx.doi.org/10.1016/j.ecolmodel.2007.04.034.

Hell, P., Nesetřil, J., 1970. Graphs and k-societies. Canad. Math. Bull. 13 (3), 375–381.

Iverson, S.J., Field, C., Don Bowen, W., Blanchard, W., 2004. Quantitative fatty acid signature analysis: a new method of estimating predator diets. Ecol. Monograph 74 (2), 211–235.

Jørgensen, S.E., Nielsen, S.N., Jørgensen, L.A., 1991. Handbook of Ecological Parameters and Ecotoxicology. Technical Report, Elsevier.

Kazanci, C., 2007. EcoNet: A new software for ecological modeling, simulation and network analysis. Ecol. Model. 208 (1), 3–8. http://dx.doi.org/10.1016/j.ecolmodel.2007.04.031.

Kazanci, C., 2020. LINX. MATLAB Central File Exchange, https://www.mathworks.com/matlabcentral/fileexchange/72143-linx.

Kazanci, C., Black, K., 2020. LINX. GitHub, https://github.com/KellyBlack/LINX.

Kemp, W., Boynton, W., 2004. Productivity, trophic structure, and energy flow in the steady-state ecosystems of Silver Springs, Florida. Ecol. Model. 178 (1–2), 43–49. http://dx.doi.org/10.1016/j.ecolmodel.2003.12.020.

Marquis, E., Niquil, N., Delmas, D., Hartmann, H., Bonnet, D., Carlotti, F., Herbland, A., Labry, C., Sautour, B., Laborde, P., et al., 2007. Inverse analysis of the planktonic food web dynamics related to phytoplankton bloom development on the continental shelf of the Bay of Biscay, French coast. Estuar. Coast. Shelf Sci. 73 (1–2), 223–235.

Morris, J.T., Christian, R.R., Ulanowicz, R.E., 2005. Analysis of size and complexity of randomly constructed food webs by information theoretic metrics. In: Belgrano, A., Scharler, U., Dunne, J., Ulanowicz, R.E. (Eds.), Aquatic Food Webs: An Ecosystem Approach. Oxford University Press, pp. 73–85.

Odum, H.T., 1957. Trophic structure and productivity of Silver Springs, Florida. Ecol. Monograph 27 (1), 55–112.

van Oevelen, D., van den Meersche, K., Meysman, F.J.R., Soetaert, K., Middelburg, J.J., Vézina, A.F., 2010. Quantifying food web flows using linear inverse models. Ecosystems 13 (1), 32–45. http://dx.doi.org/10.1007/s10021-009-9297-6.

Pacella, S.R., Lebreton, B., Richard, P., Phillips, D., DeWitt, T.H., Niquil, N., 2013. Incorporation of diet information derived from Bayesian stable isotope mixing models into mass-balanced marine ecosystem models: a case study from the Marennes-Oléron Estuary, France. Ecol. Model. 267, 127–137.

Patten, B.C., 1978. Systems approach to the concept of environment. Ohio J. Sci. 78 (4), 206–222.

Phillips, D.L., Gregg, J.W., 2003. Source partitioning using stable isotopes: coping with too many sources. Oecologia 136 (2), 261–269.

Post, D.M., 2002. Using stable isotopes to estimate trophic position: models, methods, and assumptions. Ecology 83 (3), 703–718.

Resendis-Antonio, O., 2013. Stoichiometric Matrix. In: Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H. (Eds.), Encyclopedia of Systems Biology. Springer New York, New York, NY, p. 2014.

Saint-Béat, B., Maps, F., Babin, M., 2018. Unraveling the intricate dynamics of planktonic Arctic marine food webs. A sensitivity analysis of a well-documented food web model. Prog. Oceanogr. 160, 167–185.

Savenkoff, C., Vézina, A.F., Bundy, A., 2001. Inverse Analysis of the Structure and Dynamics of the Whole Newfoundland-Labrador Shelf Ecosystem. Technical Report 2354, Canadian Technical Report of Fisheries and Aquatic Sciences 2354.

Schramski, J.R., Kazanci, C., Tollner, E.W., 2011. Network environ theory, simulation, and EcoNet®, 2.0. Environ. Model. Softw. 26 (4), 419–428. http://dx.doi.org/10.1016/j.envsoft.2010.10.003.

Sibly, R.M., Brown, J.H., Kodric-Brown, A., 2012. Metabolic Ecology: a Scaling Approach. John Wiley & Sons.

Ulanowicz, R.E., 1986. Growth and Development: Ecosystems Phenomenology. Springer.

Ulanowicz, R.E., Scharler, U.M., 2008. Least-inference methods for constructing networks of trophic flows. Ecol. Model. 210 (3), 278–286. http://dx.doi.org/10.1016/j.ecolmodel.2007.08.001.

Vézina, A.F., Pace, M.L., 1994. An inverse model analysis of planktonic food webs in experimental lakes. Can. J. Fish. Aquat. Sci. 51 (9), 2034–2044.

Vézina, A.F., Platt, T., 1988. Food web dynamics in the ocean. 1. Best-estimates of flow networks using inverse methods. Mar. Ecol. Prog. Ser. 42 (3), 269–287.

Yodzis, P., Innes, S., 1992. Body size and consumer-resource dynamics. Am. Nat. 1151–1175.