

On the interpretation of inflated correlation path weights in concentration graphs

Received: date / Accepted: date

Abstract Statistical models associated with graphs, called graphical models, have become a popular tool for representing network structures in many modern applications. Relevant features of the model are represented by vertices, edges and other higher order structures. A fundamental structural component of the network is represented by paths, which are a sequence of distinct vertices joined by a sequence of edges. The collection of all the paths joining two vertices provides a full description of the association structure between the corresponding variables. In this context, it has been shown that certain pairwise association measures can be decomposed into a sum of weights associated with each of the paths connecting the two variables. We consider a pairwise measure called an inflated correlation coefficient and investigate the properties of the corresponding path weights. We show that every inflated correlation weight can be factorized into terms, each of which is associated either to a vertex or to an edge of the path. This factorization allows one to gain insight into the role played by a path in the network by highlighting the contribution to the weight of each of the elementary units forming the path. This is of theoretical interest because, by establishing a similarity between the weights and the association measure they decompose, it provides a justification for the use of these weights. Furthermore we show how this factorization can be exploited in the computation of centrality measures and describe their use with an application to the analysis of a dietary pattern.

Keywords Betweenness centrality; Inflation factor; Partial correlation; Path weight; Undirected path.

1 Introduction

Graphical models provide a compact and efficient representation of the association structure of a multivariate distribution by means of a graph and have

Address(es) of author(s) should be given

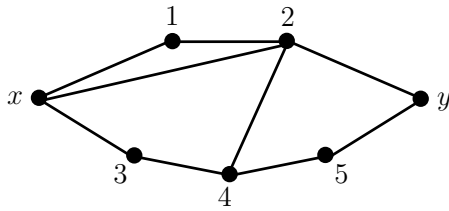


Fig. 1 Concentration graph for the covariance matrix of Table 1

become a popular tool for representing network structures in many applied contexts; see Maathuis et al. (2019) for a recent review of the state of art of graphical models. If X_V is a vector of continuous random variables then an undirected network, called a concentration graph of X_V , is constructed in such a way that every vertex is associated with a variable and a missing edge between two vertices implies that the corresponding partial correlation is equal to zero (Lauritzen, 1996). In this way, the association structure of X_V is encoded by the paths connecting the variables. Paths are the main tools used in the definition of separation criteria and therefore of the Markov properties characterizing these statistical models. More concretely, an edge joining two vertices can be regarded as a single-edge path and encodes a direct association between the corresponding variables, whereas a path made up of two or more edges represents an indirect association mediated by the intermediate variables in the path. It follows that the collection of all the paths joining a pair of vertices provides a full description of the association structure between the corresponding variables.

Example 1 The upper triangle of Table 1 contains the entries of the variance and covariance matrix of a random vector X_V with $V = \{x, 1, 2, 3, 4, 5, y\}$. From the covariance matrix one can compute the partial correlations for every pair of variables given the remaining variables (lower triangle), and Figure 1 shows a concentration graph of X_V . One can see, for instance, that the analysis of the association structure of X_x and X_y can be carried out by investigating the role played by the six different paths joining x and y in the graph; these are detailed in Table 2.

In models for directed acyclic graphs the well-established theory of path analysis (Wright, 1921) provides a method that allows one to quantify the relevance of a directed path. On the other hand, in models for undirected graphs the problem of quantifying the strength of the association encoded by paths has been investigated only more recently. Jones and West (2005) considered the measure of association between two variables provided by the covariance and showed that this quantity can be decomposed in terms of additive weights associated with the paths joining the corresponding vertices. Roverato and Castelo (2020) provided an analysis of the properties of the covariance path weights introduced by Jones and West (2005) and showed that inflation factors

Table 1 Instance of covariance matrix whose inverse is adapted to the graph of Figure 1. In this matrix the variances are all equal to 9 (main diagonal) and the covariances (upper triangle) are such that the values associated with the edges of the graph are all equal to 4.5. The lower triangle (in bold) gives the corresponding partial correlations.

	x	1	2	3	4	5	y
x	9.00	4.50	4.50	4.50	3.29	2.09	2.41
1	0.31	9.00	4.50	2.60	2.60	1.79	2.30
2	0.22	0.29	9.00	3.29	4.50	3.29	4.50
3	0.35	0.00	0.00	9.00	4.50	2.41	2.09
4	0.00	0.00	0.26	0.34	9.00	4.50	3.29
5	0.00	0.00	0.00	0.00	0.34	9.00	4.50
y	0.00	0.00	0.33	0.00	0.00	0.37	9.00

play a key role in the interpretation of these quantities; see also Roverato and Castelo (2017, 2018); Peeters et al. (2020).

The comparison of paths with different endpoint requires the use of normalized measures of association and, to this aim, Roverato and Castelo (2020) considered the weights obtained for the decomposition of correlation coefficients. Furthermore, they introduced a novel normalized measure of linear association, named the inflated correlation coefficient, and showed that the weights obtained from the decomposition of this quantity satisfy useful properties that, as far as the strength of paths is of concern, make them an appealing alternative to the classical correlation coefficients.

Here, we focus on the weights obtained from the decomposition of inflated correlations. A path weight quantifies the relevance of the corresponding path. A path can be seen as an ordered sequence of vertices and edges, and we show that every inflated correlation weight can be factorized into terms, each of which is associated either with a vertex or to an edge of the path. More specifically, every vertex is associated with an inflation factor quantifying the contribution of the variable to the path. Furthermore, every edge is associated with a partial correlation quantifying the contribution to the path of the corresponding pairwise association. This factorization allows one to gain insight into the role played by a path in the network by highlighting the contribution to the weight of each of the building blocks forming the path. This is of special interest in the comparison of paths. Moreover, it provides a theoretical justification for the use of these weights because it shows that inflated correlations can be decomposed into the sum of weights which can themselves be interpreted as inflated (partial) correlations, thereby conferring consistency between the weights and the association measure they decompose. We then show how this factorization can be used to construct betweenness centrality measures specifically designed to suit the graphical model framework. Finally, an application in the context of dietary pattern analysis is provided.

This paper is organized as follows. Background on inflation factors, inflated correlation matrices, concentration graph models and path weights is given in Section 2. In Section 3 we establish a connection between inflation factors and the determinant of inflated correlation matrices, whereas Section 4 deals with

inflated correlation weights and describes their decomposition. The betweenness centrality measures based on path weights and their decomposition are introduced in Section 5, where an application to the analysis of the eating behaviour of a group of subject is also given. Finally, Section 6 contains a brief discussion.

2 Background and notation

2.1 Inflation factors and the inflated correlation matrix

Let $X = X_V$ be a random vector indexed by a finite set $V = \{u, v, \dots, z\}$ with covariance matrix $\Sigma = \{\sigma_{uv}\}_{u,v \in V}$. We denote by $K = \{\kappa_{uv}\}_{u,v \in V}$ the concentration matrix of X_V and recall that $K = \Sigma^{-1}$. For two subsets $A, B \subseteq V$, such that $A \cap B = \emptyset$, we consider the subvectors X_A and X_B of X_V and denote by $X_A | X_B$ the residual vector deriving from the linear least square predictor of X_A on X_B (see Whittaker, 1990, p. 134). It follows that the covariance matrices of X_A and $X_A | X_B$ are Σ_{AA} and $\Sigma_{AA \cdot B} = \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA}$, respectively, where we use the convention that $\Sigma_{AA}^{-1} = (\Sigma_{AA})^{-1}$ and, similarly, $\Sigma_{AA \cdot B}^{-1} = (\Sigma_{AA \cdot B})^{-1}$. We denote by $\sigma_{uv \cdot B}$, for $u, v \in A$, the entries of $\Sigma_{AA \cdot B}$ and recall that, in the Gaussian case, $\Sigma_{AA \cdot B}$ coincides with the covariance matrix of the conditional distribution of X_A given X_B . We write \bar{A} to denote the complement of A relative to V , that is $\bar{A} = V \setminus A$ and remark that the concentration matrix of $X_A | X_{\bar{A}}$ is the submatrix of K with entries indexed by A because it follows from the rule for the inversion of a partitioned matrix that $\Sigma_{AA \cdot \bar{A}}^{-1} = K_{AA}$.

In linear regression diagnostics, the effect of multicollinearity may be quantified by means of the variance inflation factor. The inflation factor of X_v on $X_{V \setminus \{v\}}$ is defined as $\text{IF}_v = 1/(1 - \rho_{(v)(V \setminus \{v\})}^2)$ where $\rho_{(v)(V \setminus \{v\})}$ is the multiple correlation of X_v on $X_{V \setminus \{v\}}$. IF_v takes values in the interval $[1, +\infty)$; it is equal to one if and only if X_v and $X_{V \setminus \{v\}}$ are uncorrelated and its value increases as $\rho_{(v)(V \setminus \{v\})}$ increases (see Belsley et al., 2005; Chatterjee and Hadi, 2012).

Fox and Monette (1992) considered the case where one is concerned with sets of regressors rather than with individual regressors and introduced a generalized version of the variance inflation factor; specifically, for a pair of subsets $A, B \subseteq V$, with $A \cap B = \emptyset$ this is given by,

$$\text{IF}_A^B = \frac{|\Sigma_{AA}| |\Sigma_{BB}|}{|\Sigma_{A \cup B A \cup B}|}, \quad (1)$$

so that $\text{IF}_A^B = \text{IF}_v$ when $A = \{v\}$ and $B = V \setminus \{v\}$. We will refer to IF_A^B as the inflation factor of A on B and in order to simplify the notation we will write IF_A when $B = \bar{A}$. Throughout this paper, the covariance matrices we consider are assumed to be positive definite and, furthermore, we use the convention that the determinant of a submatrix whose rows and columns are indexed by

the empty set is equal to one. In this way, the inflation factor in (1) is always well-defined with $\text{IF}_A^B = 1$ whenever either $A = \emptyset$ or $B = \emptyset$.

Fox and Monette (1992) also suggested a generalization (1) to the case where X_V is partitioned into k sets, A_1, \dots, A_k . In the special case where $k = p$ so that every set contains a single variable, such inflation factor becomes a global measure of association and it is equal to $1/|\Omega|$, where $\Omega = \text{diag}(\Sigma)^{-\frac{1}{2}} \Sigma \text{diag}(\Sigma)^{-\frac{1}{2}}$ is the correlation matrix of X_V , with entries ρ_{uv} , for $u, v \in V$. This result is consistent with the usual interpretation of the determinant of Ω as common global measure of collinearity, justified by noting that $|\Omega| = 1$ for mutually uncorrelated variables and $|\Omega| = 0$ for perfectly collinear variables.

Roverato and Castelo (2020) introduced the matrix

$$\Omega^V = \text{diag}(K)^{\frac{1}{2}} \Sigma \text{diag}(K)^{\frac{1}{2}}, \tag{2}$$

and named Ω^V the inflated correlation matrix because its entries are given by $\varrho_{vv}^V = \text{IF}_v$ for every $v \in V$ and by $\varrho_{uv}^V = \rho_{uv} \sqrt{\text{IF}_u \text{IF}_v}$ for every $u, v \in V$ with $v \neq u$. Note that for $|V| = 2$ the inflated correlation ϱ_{uv}^V becomes $\varrho_{uv}^{\{u,v\}} = \rho_{uv}/(1 - \rho_{uv}^2)$. Furthermore, they showed that the determinant of Ω^V can be computed as

$$|\Omega^V| = \frac{|\Sigma|}{\prod_{v \in V} \sigma_{vv \cdot V \setminus \{v\}}}$$

and that this determinant provides an alternative global measures of linear association which, like $1/|\Omega|$, takes values in the interval $[1, +\infty)$ and is equal to one if and only if Σ is diagonal.

The quantities defined in this section can also be computed with respect to the distribution of $X_A \mid X_B$. More specifically, we will denote by $\Omega_{AA \cdot B}^V = \{\varrho_{uv \cdot B}^V\}_{u,v \in A}$ the inflated correlation matrix of $X_A \mid X_B$ and we remark that if $A \cup B = V$ then, similarly to the covariance matrix $\Sigma_{AA \cdot \bar{A}}$, the matrix $\Omega_{AA \cdot B}^V$ can be computed as $\Omega_{AA \cdot B}^V = \Omega_{AA}^V - \Omega_{AB}^V (\Omega_{BB}^V)^{-1} \Omega_{BA}^V$.

2.2 Concentration graph models

An undirected graph with vertex set V is a pair $\mathcal{G} = (V, \mathcal{E})$ where \mathcal{E} is a set of edges, which are unordered pairs of vertices; formally $\mathcal{E} \subseteq V \times V$. The graphs we consider have no self-loops, that is $\{v, v\} \notin \mathcal{E}$ for any $v \in V$. A path of length $k \geq 2$ between x and y in \mathcal{G} is a sequence $\pi = \langle x = v_1, \dots, v_k = y \rangle$ of distinct vertices such that $\{v_i, v_{i+1}\} \in \mathcal{E}$ for every $i = 1, \dots, k - 1$. We denote by $V(\pi) \subseteq V$ and $\mathcal{E}(\pi) \subseteq \mathcal{E}$ the set of vertices and edges of the path π , respectively. We write π_{xy} when we want to make more explicit which are the endpoints of the path and, furthermore, when clear from the context we will set $P \equiv V(\pi)$ thereby improving the readability of sub- and super-scripts. For a pair of vertices $x, y \in A$ we denote by $\Pi_{xy}^V \equiv \Pi_{xy}$ the collection of all paths between x and y in \mathcal{G} .

If K is the concentration matrix of X_V then for every $u, v \in V$ it holds that

$$\sigma_{vv \cdot V \setminus \{v\}} = \frac{1}{\kappa_{vv}} \quad \text{and} \quad \rho_{uv|V \setminus \{u,v\}} = \frac{-\kappa_{uv}}{\sqrt{\kappa_{uu}\kappa_{vv}}}$$

where $\rho_{uv|V \setminus \{u,v\}}$ is the partial correlation coefficient of X_u and X_v given $X_{V \setminus \{u,v\}}$; see Whittaker (1990, section 5.7). It follows that $\rho_{uv|V \setminus \{u,v\}} = 0$ if and only if $\kappa_{uv} = 0$ and we say that K is adapted to a graph $\mathcal{G} = (V, \mathcal{E})$ if for every $\kappa_{uv} \neq 0$, with $u \neq v$, it holds that $\{u, v\} \in \mathcal{E}$ and, accordingly, we call \mathcal{G} a concentration graph of X_V .

The concentration graph model (Cox and Wermuth, 1996) with graph $\mathcal{G} = (V, \mathcal{E})$ is the family of multivariate normal distributions whose concentration matrix is adapted to \mathcal{G} . The latter model has also been called a covariance selection model (Dempster, 1972) and a graphical Gaussian model (Whittaker, 1990); we refer the reader to Lauritzen (1996) for details and discussion.

2.3 Decomposition of association measures over \mathcal{G}

In the analysis of concentration graph models, Jones and West (2005) showed that the covariance between two variables can be computed as the sum of weights associated with the paths joining the two variables. More specifically, if the concentration matrix K of X_V is adapted to the graph $\mathcal{G} = (V, \mathcal{E})$ then for every $x, y \in V$ it holds that

$$\sigma_{xy} = \sum_{\pi \in \Pi_{xy}} \omega(\pi, \Sigma) \quad (3)$$

where

$$\omega(\pi, \Sigma) = (-1)^{|\pi|+1} \frac{|K_{\bar{P}\bar{P}}|}{|K|} \prod_{\{u,v\} \in \mathcal{E}(\pi)} \kappa_{uv}. \quad (4)$$

The quantity $\omega(\pi, \Sigma)$ in (4) represents the contribution of the path π to the covariance σ_{xy} and for this reason we call it the covariance weight of π relative to X_V . More generally, we will refer to (3) with the name of the covariance decomposition over \mathcal{G} .

An issue concerning the covariance decomposition in (3) is the interpretation of the values taken by the weights of a path. From this perspective, Roverato and Castelo (2020) showed that every covariance weight can be factorized as $\omega(\pi, \Sigma) = \omega(\pi, \Sigma_{P\bar{P}}) \times \text{IF}_P$, with the two factors which provide two clearly distinct pieces of information. More specifically, the first term, $\omega(\pi, \Sigma_{P\bar{P}})$, is the covariance weight computed on the distribution of $X_P | X_{\bar{P}}$ and captures the strength of the path, after adjusting for all the variables outside the path, while the inflation factor IF_P captures the connectivity of the vertices of the path with the rest of the multivariate system.

Table 2 All the paths between x and y in the graph of Figure 1 with the corresponding weights, proportion of covariance due to the weight, partial weight and inflation factor, where $P_i = V(\pi_i)$.

i	π_i	$\omega(\pi_i, \Sigma)$	perc.	$\omega(\pi_i, \Sigma_{P_i, P_i, \bar{P}_i})$	IF $_{P_i}$
1	$\langle x, 1, 2, y \rangle$	0.487	20.19 %	0.249	1.956
2	$\langle x, 1, 2, 4, 5, y \rangle$	0.062	2.58 %	0.039	1.577
3	$\langle x, 2, y \rangle$	1.200	49.75 %	0.462	2.598
4	$\langle x, 2, 4, 5, y \rangle$	0.153	6.35 %	0.071	2.169
5	$\langle x, 3, 4, 2, y \rangle$	0.215	8.93 %	0.095	2.272
6	$\langle x, 3, 4, 5, y \rangle$	0.294	12.20 %	0.144	2.049

Example 2 Table 1 gives the entries of a covariance matrix, Σ , whose inverse K is adapted to the graph \mathcal{G} of Figure 1. There are $|I_{xy}| = 6$ paths between x and y in \mathcal{G} and these are given in Table 2, together with the corresponding weights. The covariance of X_x and X_y is equal to $\sigma_{xy} = 2.411$, and it can be readily checked that $\sum_{\pi \in I_{xy}} \omega(\pi, \Sigma) = \sum_{i=1}^6 \omega(\pi_i, \Sigma) = \sigma_{xy}$. Because the six weights have the same sign (they are actually all positive) it makes sense to include in Table 2 a column with the relative contribution of every path to the covariance. This shows, for example, that almost 50% of the value of σ_{xy} is due to path π_3 . The decomposition of covariance weights into partial weights and inflation factor is given in the last two columns of Table 2. One can see that the relevance of path π_3 with respect to the other paths is mainly due to its partial weight because its inflation factor is only slightly larger than those of the other paths.

It can be shown that the value of a covariance weight depends on the scale of the variables which are endpoint of the path. Hence, in order to compare paths with different endpoints it is necessary to deal with normalized quantities. Roverato and Castelo (2020) noticed that the decomposition in (3) is not restricted to covariance matrices but it can be straightforwardly extended to any positive definite matrix $\Gamma = \{\gamma_{uv}\}_{u,v \in V}$ obtained as $\Gamma = \Delta \Sigma \Delta$ where $\Delta = \{\delta_{uv}\}_{u,v \in V}$ is a diagonal matrix with nonzero diagonal entries. More specifically, both Ω and Ω^V are specific instances of this general setting with $\Delta = \text{diag}(\Sigma)^{-\frac{1}{2}}$ and $\Delta = \text{diag}(K)^{\frac{1}{2}}$, respectively. Indeed, both correlations and inflated correlations are normalized measures of association and for this reason the corresponding weights are of interest. More specifically, Roverato and Castelo (2020) provided the following decomposition of inflated correlations,

$$\varrho_{xy}^V = \sum_{\pi \in I_{xy}} \omega(\pi, \Omega^V) \quad (5)$$

where

$$\omega(\pi, \Omega^V) = |\Omega_{PP}^V| \prod_{\{u,v\} \in \mathcal{E}(\pi)} \rho_{uv|V \setminus \{u,v\}}; \quad (6)$$

we refer to Roverato and Castelo (2020) for details on the properties of inflated correlation weights. Here we remark that, as well as the weight in (4)

also inflated correlation weights can be factorized into a partial weight and an inflation factor, $\omega(\pi, \Omega^V) = \omega(\pi, \Omega_{PP.\bar{P}}^V) \times \text{IF}_P$ and, furthermore, all the factors in (6) are feasible of a clear interpretation because it is the product of the partial correlations corresponding to the edges of the path and of $|\Omega_{PP}^V| \geq 1$ which “inflates” the partial correlations.

3 On the relationship between inflation factors and inflated correlation matrices

In the theory of path weights a key role is played by both inflation factors and inflated correlation matrices. The inflation factor IF_A^B provides a well-established way to quantify the linear association between two random vectors X_A and X_B . On the other hand, the inflated correlation matrix Ω^V was firstly introduced by Roverato and Castelo (2020) who showed that the determinant of this matrix can be regarded as multivariate generalization of the inflation factor, and therefore as a global measure of linear association of X_V . As well as the inflation factor, also $|\Omega^V|$ takes values in the interval $[1; +\infty)$ where the value 1 represents absence of linear association, and in this section we formally establish a connection between these two quantities. The stated relationship between inflation factors and inflated correlation matrices will be exploited in the next section to the computation and interpretation of path weights. However, it is also of theoretical interest because it provides a clear way to interpret the value of $|\Omega^V|$, thereby allowing us to gain insight into the type of information conveyed by this quantity.

We first need to prove the following lemma.

Lemma 1 *Let X_V be a random vector indexed by a finite set V and let Ω^V be its inflated correlation matrix. If $A, B \subseteq V$ is a pair of subsets of V with $A \neq \emptyset$, then for every $v \in A$ the following factorizations hold true,*

- (i) $\text{IF}_A^B = \text{IF}_{v|A'}^B \times \text{IF}_{A'}^B$;
- (ii) $|\Omega_{AA}^V| = |\Omega_{A'A'.\{v\}}^V| \times \text{IF}_v$;
- (iii) $|\Omega_{AA.B}^V| = |\Omega_{A'A'.B \cup \{v\}}^V| \times \text{IF}_{v|B}$.

where $A' = A \setminus \{v\}$. Note that in the case where $A = \{v\}$ and $B = V \setminus \{v\}$ the identity (iii) becomes $|\Omega_{\{v\}\{v\}.V \setminus \{v\}}^V| = \Omega_{\{v\}\{v\}.V \setminus \{v\}}^V = 1$.

Proof In the case where $A = \{v\}$, so that $A' = \emptyset$, the results (i) and (ii) are trivially true because the identity (i) becomes $\text{IF}_A^B = \text{IF}_{v|\emptyset}^B$ whereas the identity (ii) is $|\Omega_{AA}^V| = \Omega_{\{v\}\{v\}}^V = \text{IF}_v$. Furthermore, in (iii) $|\Omega_{AA.\bar{A}}^V| = \Omega_{\{v\}\{v\}.\bar{A}}^V = \sigma_{vv.\bar{A}} \times \kappa_{vv} = \frac{\sigma_{vv.\bar{A}}}{\sigma_{vv.\bar{A}}} = 1$.

The identity (i) can be shown by using the alternative formulation of the inflation factor given in Roverato and Castelo (2020, eqn. (4)) and then applying the Schur’s determinant identity formula as follows,

$$\text{IF}_A^B = \frac{|\Sigma_{AA}|}{|\Sigma_{AA.B}|} = \frac{\sigma_{vv.A'}}{\sigma_{vv.A' \cup B}} \times \frac{|\Sigma_{A'A'}|}{|\Sigma_{A'A'.B}|} = \text{IF}_{v|A'}^B \times \text{IF}_{A'}^B; \quad (7)$$

notice that (7) is still valid if either A' or B are equal to the empty set because if $A' = \emptyset$ then, by convention, $|\Sigma_{A'A'}| = |\Sigma_{A'A'.B}| = \mathbf{IF}_{A'}^B$ and, similarly, if $B = \emptyset$ then $\mathbf{IF}_{v|A'}^B = \mathbf{IF}_{A'}^B = 1$.

In order to show (iii) we first apply the Schur's determinant identity formula to $|\Omega_{AA.B}^V|$ as follows,

$$|\Omega_{AA.B}^V| = |\Omega_{A'A'.B \cup \{v\}}^V| \times |\Omega_{\{v\}\{v\}.B}^V|$$

with

$$|\Omega_{\{v\}\{v\}.B}^V| = \Omega_{\{v\}\{v\}.B}^V = \kappa_{vv} \times \sigma_{vv.B} = \frac{\sigma_{vv.B}}{\sigma_{vv.V \setminus \{v\}}} = \mathbf{IF}_{v|B}. \quad (8)$$

where κ_{vv} is the relevant entry of the concentration matrix K of X_V . We also notice that in the case where $A = \{v\}$ and $B = \bar{A} = V \setminus \{v\}$ then (8) becomes

$$|\Omega_{\{v\}\{v\}.V \setminus \{v\}}^V| = \Omega_{\{v\}\{v\}.V \setminus \{v\}}^V = 1.$$

Finally, (ii) is a special case of (iii) obtained by setting $B = \emptyset$.

It is worth remarking that, in order to apply Lemma 1 in the case where either A' or B are empty, one has to recall that in this paper we use the convention that the determinant of matrices indexed by the empty set are equal to one and this also implies, for instance, that $\mathbf{IF}_{\emptyset}^B = \mathbf{IF}_v^0 = 1$.

For a nonempty set $A \subseteq V$ consider a numbering of the elements of $A = \{v_1, \dots, v_{|A|}\}$. Then the predecessors of a vertex $v_i \in A$, denoted by $\text{pr}(v_i)$, are those vertices that have lower number than v_i , formally $\text{pr}(v_i) = \{v_1, \dots, v_{i-1}\}$; hence, $\text{pr}(v_1) = \emptyset$.

Theorem 2 *Let X_V be a random vector indexed by a finite set V with $|V| = p$, and let Ω^V be its inflated correlation matrix. Then for any nonempty subset $A \subseteq V$ and any numbering of the elements of $A = \{v_1, \dots, v_{|A|}\}$ it holds that,*

$$|\Omega_{AA}^V| = \prod_{i=1}^{|A|} \mathbf{IF}_{v_i|\text{pr}(v_i)}. \quad (9)$$

Furthermore, when $A = V$ the last term of the factorization (11) is equal to one, that is $\mathbf{IF}_{v_p|\text{pr}(v_p)} = 1$.

Proof If $|A| = 1$ then (9) is equivalent to point (ii) of Lemma 1. Hence we assume $|A| = q \geq 2$ and consider an arbitrary numbering $A = \{v_1, \dots, v_q\}$ of the elements of A . Hence, we can first apply the factorization (ii) of Lemma 1 to obtain $|\Omega_{AA}^V| = |\Omega_{A'A'.\{v_1\}}^V| \times \mathbf{IF}_{v_1}$, where $A' = A \setminus \{v_1\}$ and then we can apply (iii) iteratively to v_2, \dots, v_{q-1} to obtain the factorization in (9).

This theorem deals with an arbitrary submatrix of Ω^V and shows that its determinant can be written as the product of inflation factors. On the right side of (9) the elements of A are taken one at the time and the term relative to v_i , for $i = 1, \dots, |A|$, captures the additional contribution of X_{v_i}

to $|\Omega_{AA}^V|$ with respect to the previous variables considered. Concretely, this is given by $\text{IF}_{v_i|\text{pr}(v_i)}$, the inflation factor of X_{v_i} computed on the distribution of $X_{V \setminus \text{pr}(v_i)} \mid X_{\text{pr}(i)}$, that is the inflation factor of X_{v_i} on all the remaining variables linearly adjusted for $X_{\text{pr}(v_i)}$. It is worth noting that the contribution of the last variable in the numbering is $\text{IF}_{v_{|A|}|\text{pr}(v_{|A|})}$, that is the inflation factor of $X_{v_{|A|}}$ on $X_{\bar{A}}$ adjusted for $X_{A \setminus \{v_{|A|}\}}$, and in the case where $A = V$ this is equal to 1 and therefore uninfluential. In turn, this implies that $|\Omega^V| = |\Omega_{V \setminus \{v\}}^V|$ for every $v \in V$. It is also useful to compare (9) with the following factorization of IF_A .

Theorem 3 *Let X_V be a random vector indexed by a finite set V . Then for any nonempty subset $A \subseteq V$ and any numbering of the elements of $A = \{v_1, \dots, v_{|A|}\}$ it holds that,*

$$\text{IF}_A = \prod_{i=1}^{|A|} \text{IF}_{v_i|\text{pr}(v_i)}^{\bar{A}}. \quad (10)$$

Proof The result follows from the iterative application of the factorization (i) of Lemma 1.

In a similar fashion to (9), each term of the factorization (10) captures the additional contribution of X_{v_i} to IF_A . In order to understand the different type of information provided by $|\Omega_{AA}^V|$ with respect to IF_A it is useful to compare every term on the right hand side of (9) with the corresponding term in (10). In this way we see that both $\text{IF}_{v_i|\text{pr}(v_i)}^{\bar{A}}$ and $\text{IF}_{v_i|\text{pr}(v_i)}$ are computed on the distribution of $X_{V \setminus \text{pr}(v_i)} \mid X_{\text{pr}(i)}$, however the former inflation factor only involves the linear association between X_{v_i} and the variables not in A , $X_{\bar{A}}$, whereas the latter inflation factor involves the linear association between X_{v_i} and both the variables not in A and the remaining variables in A , that is both $X_{\bar{A}}$ and $X_{A \setminus \text{pr}(v_i) \cup \{v_i\}}$. The following result gives an additional relationship between IF_A and $|\Omega_{AA}^V|$.

Corollary 4 *Let X_V be a random vector indexed by a finite set V and let Ω^V be its inflated correlation matrix. If $A, B \subseteq V$ is a pair of subsets of V with $A \neq \emptyset$, then for any numbering of the elements of $A = \{v_1, \dots, v_{|A|}\}$ it holds that,*

$$|\Omega_{AA \cdot B}^V| = \prod_{i=1}^{|A|} \text{IF}_{v_i|B \cup \text{pr}(v_i)}. \quad (11)$$

Furthermore, when $A \cup B = V$ so $B = \bar{A}$ the last term of the factorization (11) is equal to one, that is $\text{IF}_{v_{|A|}|\bar{A} \cup \text{pr}(v_{|A|})} = \text{IF}_{v_{|A|}|V \setminus \{v\}} = 1$ and, furthermore it holds that

$$|\Omega_{AA}^V| = \text{IF}_A \times |\Omega_{AA \cdot \bar{A}}^V|. \quad (12)$$

Proof The equality (11) follows from the iterative application of the factorization in (iii) of Lemma 1. The equality (12) follows by noticing that $\text{IF}_{v_i|\text{pr}(v_i)} / \text{IF}_{v_i|\bar{A} \cup \text{pr}(v_i)} = \text{IF}_{v_i|\text{pr}(v_i)}^{\bar{A}}$, for every $i = 1 \dots, |A|$, so that it follows from (9) and (11) that $|\Omega_{AA}^V| / |\Omega_{AA \cdot \bar{A}}^V| = \text{IF}_A$ in (10)

Equation (12) show that $|\Omega_{AA}^V|$ can be computed as the product of two quantities, $|\Omega_{AA\bar{A}}^V|$ and IF_A . The former, $|\Omega_{AA\bar{A}}^V|$, is a measure of global association of variables in X_A linearly adjusted for $X_{\bar{A}}$ whereas the former, IF_A , measures the strength of the linear association between X_A and $X_{\bar{A}}$. Recall that both $|\Omega_{AA\bar{A}}^V| \geq 1$ and $\text{IF}_A \geq 1$ and therefore $|\Omega_{AA}^V| = 1$ if and only if $\Sigma_{AA\bar{A}}$ is diagonal and $\Sigma_{A\bar{A}} = 0$.

4 Decomposition of inflated correlation weights

In this section we consider the inflated correlation weights in (6) and exploit the results of the previous section to provide an alternative formulation of these quantities that identifies the role played by every vertex and edge of the path.

Assume that the concentration matrix K of the random vector X_V is adapted to the undirected graph $\mathcal{G} = (V, \mathcal{E})$ and let $\pi = \langle v_1, \dots, v_k \rangle$ be a path between v_1 and v_k in \mathcal{G} . The vertices $P = V(\pi)$ of π are naturally ordered along the path and, more precisely, because the paths we consider are undirected, every path identifies two different orderings of its vertices each starting from one of the two endpoints of the path. We will refer to these orderings as the two natural numberings of the vertices of the path.

Proposition 5 *Let K be the concentration matrix of X_V . If K is adapted to the graph $\mathcal{G} = (V, \mathcal{E})$ then for every path $\pi = \langle v_1, \dots, v_k \rangle$ between v_1 and v_k in \mathcal{G} it holds that*

$$\omega(\pi, \Omega^V) = \prod_{i=1}^{|\mathcal{V}(\pi)|} \text{IF}_{v_i | \text{pr}(v_i)} \prod_{\{u,v\} \in \mathcal{E}(\pi)} \rho_{uv | \mathcal{V} \setminus \{u,v\}}. \quad (13)$$

Proof The result follows from the application of Theorem 2 to the definition of $\omega(\pi, \Omega^V)$ in (6).

We illustrate the application of Proposition 5 with an example.

Example 3 The covariance matrix given in the Example 1 can be inverted to obtain a concentration matrix that is adapted to the graph depicted in Figure 1. The path $\pi_{xy} = \langle x, 1, 2, y \rangle$ has inflated correlation weight equal to $\omega(\pi_{xy}, \Omega^V) = 0.09$ and if we apply Proposition 5 with respect to the natural vertex numbering starting from the endpoint x we can associate to every vertex of the path an inflation factor and to every edge a partial correlation, as follows, where we write $\rho_{uv | rest}$ to denote the partial correlation between X_u and X_v given all the remaining variables $X_{V \setminus \{u,v\}}$.

x	—————	1	—————	2	—————	y
IF_x	$\rho_{x1 rest}$	$\text{IF}_{1 x}$	$\rho_{12 rest}$	$\text{IF}_{2 x1}$	$\rho_{2y rest}$	$\text{IF}_{y x12}$
1.74	0.31	1.13	0.29	1.33	0.33	1.18

The factorization of $\omega(\pi, \Omega^V)$ in (13) can be carried out with respect to any of the two natural numberings of the vertices of the path.

Example 3 (Continued) An alternative decomposition of the weight $\omega(\pi_{xy}, \Omega^V)$ for $\pi_{xy} = \langle x, 1, 2, y \rangle$ can be obtained from the natural ordering of the vertices of the path starting from the endpoint y , as follows.

$$\begin{array}{ccccccc}
 x & \text{-----} & 1 & \text{-----} & 2 & \text{-----} & y \\
 \text{IF}_{x|12y} & \rho_{x1|rest} & \text{IF}_{1|2y} & \rho_{12|rest} & \text{IF}_{2|y} & \rho_{2y|rest} & \text{IF}_y \\
 1.16 & 0.31 & 1.12 & 0.29 & 1.49 & 0.33 & 1.58
 \end{array}$$

The possible choice of different vertex numbering may be an advantage. For instance, as shown below, the comparison of the two paths π_{xy} and π_{xz} in (16) becomes straightforward if one considers for both weights the natural numbering starting from the endpoint x of the two paths. On the other hand, the paths we consider are undirected and it is desirable to have a decomposition of path weights that is symmetric with respect to the two endpoints of the path. To this aim, for a path π_{xy} we consider the two natural ordering of its vertices and denote by $\text{pr}_x(v)$ and $\text{pr}_y(v)$ the predecessor of $v \in V$ with respect to numbering starting from x and y respectively. Hence, we introduce an inflation factor computed as the geometric mean of the corresponding inflation factors in the two natural numbering of the vertices.

$$\text{IF}_{\langle v, \pi_{xy} \rangle} = \left(\text{IF}_{v|\text{pr}_x(v)} \times \text{IF}_{v|\text{pr}_y(v)} \right)^{\frac{1}{2}} \quad (14)$$

and we will simply write $\text{IF}_{\langle v \rangle}$ when it is clear from the context which path we are referring to. We can now state the main result of this section.

Theorem 6 *Let K be the concentration matrix of X_V . If K is adapted to the graph $\mathcal{G} = (V, \mathcal{E})$ then for every path $\pi = \langle v_1, \dots, v_k \rangle$ between v_1 and v_k in \mathcal{G} it holds that*

$$\omega(\pi, \Omega^V) = \prod_{v \in V(\pi)} \text{IF}_{\langle v \rangle} \prod_{\{u, v\} \in \mathcal{E}(\pi)} \rho_{uv|V \setminus \{u, v\}}. \quad (15)$$

Proof The result follows because

$$\begin{aligned}
 \prod_{v \in V(\pi_{xy})} \text{IF}_{\langle v, \pi_{xy} \rangle} &= \prod_{v \in V(\pi_{xy})} \left(\text{IF}_{v|\text{pr}_x(v)} \times \text{IF}_{v|\text{pr}_y(v)} \right)^{\frac{1}{2}} \\
 &= \left(\prod_{v \in V(\pi_{xy})} \text{IF}_{v|\text{pr}_x(v)} \right)^{\frac{1}{2}} \left(\prod_{v \in V(\pi_{xy})} \text{IF}_{v|\text{pr}_y(v)} \right)^{\frac{1}{2}} \\
 &= |\Omega_{PP}^V|.
 \end{aligned}$$

The decomposition of $\omega(\pi, \Omega^V)$ in (15) is uniquely associated to a path and can effectively capture the role played by the building blocks of the path, as shown in the example below.

Example 3 (Continued) The decomposition (15) of $\omega(\pi_{xy}, \Omega^V)$ is as follows.

x	$\rho_{x1 rest}$	$\text{IF}_{\langle 1 \rangle}$	$\rho_{12 rest}$	$\text{IF}_{\langle 2 \rangle}$	$\rho_{2y rest}$	$\text{IF}_{\langle y \rangle}$
1.42	0.31	1.12	0.29	1.41	0.33	1.37

Unlike each of the two decompositions obtained from the two natural numbering of vertices, this decomposition shows that the variables X_x , X_2 and X_y play a similar role in the path. On the other and, the smallest inflation factor is associated with the vertex 1 and, interestingly, this is the only vertex in the path that is not linked with any vertex outside the path.

In graphical modelling the distinction between directed and undirected edges is important. A directed edge indicates the direction of dependence of a response on an explanatory variable. In a directed path every intermediate vertex is at the same time a response for the previous variables and explanatory for the following variables. Thus, for any directed graph there exists a natural ordering of variables that can be exploited to obtain a recursive factorization of the probability distribution. In turn, the terms of such factorization can be used to assess the contribution of each of the elementary units forming the path. On the other hand, undirected edges represent symmetric relationships whose interpretation is less straightforward, possibly resulting from a feedback relationship (Lauritzen and Richardson, 2002). Thus, when investigating the interpretation of a path weight, the two endpoints of the undirected path need to be put on an equal footing. The decomposition given Theorem 6 satisfies this requirement because it is obtained from the geometric mean of the two alternative decompositions of the same weight with respect to the two natural orderings of the vertices. From this viewpoint, Proposition 5 could have been stated as a lemma preliminary to Theorem 6. However, we deem that Proposition 5 has its own interest because it can be readily applied to the comparison of paths. Consider, for instance, the case where we have a path $\pi_{xy} = \langle x, \dots, y \rangle$ and $\pi_{xz} = \langle \pi_{xy}, z \rangle = \langle x, \dots, y, z \rangle$ so that π_{xz} is exactly one edge longer than π_{xy} .

$$\begin{aligned}
 \pi_{xz} : & \quad x \text{ --- } \dots \text{ --- } y \text{ --- } z & (16) \\
 \pi_{xy} : & \quad x \text{ --- } \dots \text{ --- } y
 \end{aligned}$$

Then we can compute the ratio of the two relevant weights thereby obtaining,

$$\frac{\omega(\pi_{xz}, \Omega^V)}{\omega(\pi_{xy}, \Omega^V)} = \rho_{yz|rest} \times \text{IF}_{z|V(\pi_{xy})}. \tag{17}$$

The path π_{xz} has one edge and one vertex more than π_{xy} , and the contribution of these additional components can be quantified as the product of the partial correlation associated with the additional edge and the inflated correlation associated with the additional vertex. Although the role played by the partial correlation in (17) is somehow intuitive because in concentration graph models partial correlations are naturally associated with edges, the role played by the inflation factor is more subtle. The relevance of a path within a network also depends on how its vertices interact with the rest of the network. The inflation factor in (17) quantifies the contribution of the additional variable X_z to the interaction with the rest of the network. This quantity is computed after the variables are adjusted with respect to $X_{V(\pi_{xy})}$, so as that it gives the “additional” contribution of z with respect to the vertices already present in $V(\pi_{xy})$. In fact, if the additional vertex z is connected with vertices forming the path but with no other vertex outside $V(\pi_{xy})$, then $\text{IF}_{z|V(\pi_{xy})} = 1$.

We close this section by remarking that the factorization in (15) is also of theoretical interest. Equation (5) shows the decomposition of the inflated correlation ϱ_{xy}^V over the paths of \mathcal{G} where $\varrho_{xy}^V = \rho_{xy}(\text{IF}_x \times \text{IF}_y)^{\frac{1}{2}}$. It is theoretically relevant that an association measure might be decomposed into path weights with have the same type of interpretation. An inflated correlation coefficients is obtained from the product of a correlation and the geometric mean of two inflation factor. The right hand side of (15) is consistent with this type of interpretation because its elements are (partial) correlations and quantities obtained as geometric mean of two inflation factors.

5 Application to the construction of betweenness centrality measures

In these section we apply Theorem 6 to the construction of centrality measures and describe their use to the analysis of a network representing the eating behaviour of a group of subjects.

Undirected graphs can effectively be used to model the structure of complex systems and, in many applied contexts, the association network is expected to be very heterogeneous with some vertices and edges being more important than others in some sense. This importance can be referred to as network centrality and it is typically quantified by means of centrality measures; see Rodrigues (2019). Centrality is one of the most fundamental metrics in network science, but there is no general definition of centrality and a wide range of centrality measures focusing on different features of the network are available. One of the most prominent measure of centrality, called betweenness centrality, relies on the idea that information flows along paths. The most widely used betweenness measure is due to Freeman (1977) and it is based on the idea that a vertex has a high betweenness centrality if a large number of shortest paths crosses it. Accordingly, betweenness of a vertex is computed by summing up the fractions of shortest paths between every pairs of vertices that pass through it.

The choice to focus on shortest paths was motivated in a context of social network analysis. On the other hand, in other fields of application the assumption that information flows only along shortest paths is not justified. This has led to the introduction of alternative betweenness centrality measures where all paths contribute, possibly with different values, to the computation (Freeman et al., 1991; Newman, 2005). More specifically, one can use different criteria to quantify the relevance of a path to the centrality of a vertex and this results in different betweenness measures. From this perspective, we consider the following comprehensive way to compute the betweenness of the vertex $v \in V$,

$$B^X(v) = \sum_{x,y \in V \setminus \{v\}} B_{xy}^X(v) \quad (18)$$

where $B_{xy}^X(v)$ is a measure of betweenness of vertex v relative to vertices x and y , $x \neq y$, based on the criterion χ . Although centrality is most commonly computed for vertices, also edge centrality is of interest; see Girvan and Newman (2002); Bröhl and Lehnertz (2019) and references therein. Hence, similarly to (18), we compute betweenness of an edge $\{u, v\}$ according to the criterion χ , as

$$B^X(\{u, v\}) = \sum_{x,y \in V} B_{xy}^X(\{u, v\}). \quad (19)$$

Some applications of betweenness centrality in graphical models can be found in Bringmann et al. (2019), Dablander and Hinne (2019), Peeters et al. (2020) and Roverato and Castelo (2020). However, the construction of centrality measures specifically designed to suit the graphical model framework is a recent, and largely unexplored, area of research. In the following, we consider three different types of vertex/edge betweenness centrality. The first type is based exclusively on the graph structure and therefore not specific of the graphical model field. The second and third types are specific of concentration graph models and are based on the theory of path weights and on the weight decomposition given in Section 4, respectively.

We refer to the first centrality measure with the name *basic* because it differs from that of Freeman et al. (1991) only from the fact that it is computed using all paths rather than shortest paths. More formally, it is denoted by $B^\phi(\cdot)$ and it is obtained by applying

$$B_{xy}^\phi(v) = \frac{\sum_{\pi \in \Pi_{xy}} I_v(\pi)}{|\Pi_{xy}|} \quad \text{and} \quad B_{xy}^\phi(\{u, v\}) = \frac{\sum_{\pi \in \Pi_{xy}} I_{\{u, v\}}(\pi)}{|\Pi_{xy}|}, \quad (20)$$

in (18) and (19), respectively. Here, $I_v(\pi)$ denotes the indicator function that takes value one if $v \in V(\pi)$ and zero otherwise; similarly, $I_{\{u, v\}}(\pi) = 1$ if $\{u, v\} \in \mathcal{E}(\pi)$ and zero otherwise.

We now turn to the specific case of concentration graph models. This is done by keeping into account the meaning and role that paths play in these

models, and we deem that the theory of path weights provides a natural framework to address this issue. Consider the criterion $\chi = \omega$ such that

$$B_{xy}^\omega(v) = \frac{\sum_{\pi \in \Pi_{xy}} |\omega(\pi, \Omega^V)| I_v(\pi)}{\sum_{\pi \in \Pi_{xy}} |\omega(\pi, \Omega^V)|} \quad (21)$$

and

$$B_{xy}^\omega(\{u, v\}) = \frac{\sum_{\pi \in \Pi_{xy}} |\omega(\pi, \Omega^V)| I_{\{u, v\}}(\pi)}{\sum_{\pi \in \Pi_{xy}} |\omega(\pi, \Omega^V)|}, \quad (22)$$

so that every path contributes to the computation with its absolute inflated correlation weight. The vertex betweenness centrality based on (21) was first introduced by Roverato and Castelo (2020) whereas in (22) we use the same criterion to introduce a novel edge betweenness centrality based on path weights. Note that, if all the paths between x and y have the same sign, then $B_{xy}^\omega(\cdot)$ can be interpreted as the proportion of the inflated correlation coefficients between X_x and X_y due to the paths involving the relevant vertex/edge. It is also worth remarking that, in fact, $B_{xy}^\omega(v)$ can be equally interpreted as the proportion of covariance or correlation. Hereafter, we will refer to $B^\omega(\cdot)$ as to the *weight* betweenness.

The criterion applied in (21) and (22) is, perhaps, the most straightforward way to apply the theory of path weights in the computation of betweenness centralities. A more subtle way may be obtained by considering the factorization in Theorem 6 and assigning to every path a value reflecting the role played by the relevant vertex/edge in the determination of the path weight. More specifically, we define

$$B_{xy}^\nu(v) = \sum_{\pi \in \Pi_{xy}} (\text{IF}_{\langle v, \pi_{xy} \rangle} - 1) I_v(\pi) \quad (23)$$

where $\text{IF}_{\langle v, \pi_{xy} \rangle}$, given in (14), represents to the contribution of vertex v to the path π . More specifically, $\text{IF}_{\langle v, \pi_{xy} \rangle} \geq 1$ ‘‘inflates’’ the path $\omega(\pi, \Omega^V)$ of a factor equal to $(\text{IF}_{\langle v, \pi_{xy} \rangle} - 1)$. Similarly, from Theorem 6, the contribution of an edge $\{u, v\}$ to the weight of a path π between x and y may be quantified by $(\text{IF}_{\langle u, \pi_{xy} \rangle} |\rho_{uv|rest}| \text{IF}_{\langle v, \pi_{xy} \rangle} - 1)$, thereby giving,

$$B_{xy}^\nu(\{u, v\}) = \sum_{\pi \in \Pi_{xy}} (\text{IF}_{\langle u, \pi_{xy} \rangle} |\rho_{uv|rest}| \text{IF}_{\langle v, \pi_{xy} \rangle} - 1) I_{\{u, v\}}(\pi). \quad (24)$$

We will refer to $B^\nu(\cdot)$ as to the *inflation* betweenness.

The rest of this section is devoted to an application where we compare the behaviour of the three types of centrality measures on a food network. Hoang et al. (2020) applied concentration graph models to learn the networks describing the eating behaviour of some distinct groups of subjects. Here, we focus on the network, given in Figure 2, which represents the main dietary pattern for the group of men. Every vertex is associated with a food group whereas edges show how food groups are consumed in relation to each other.

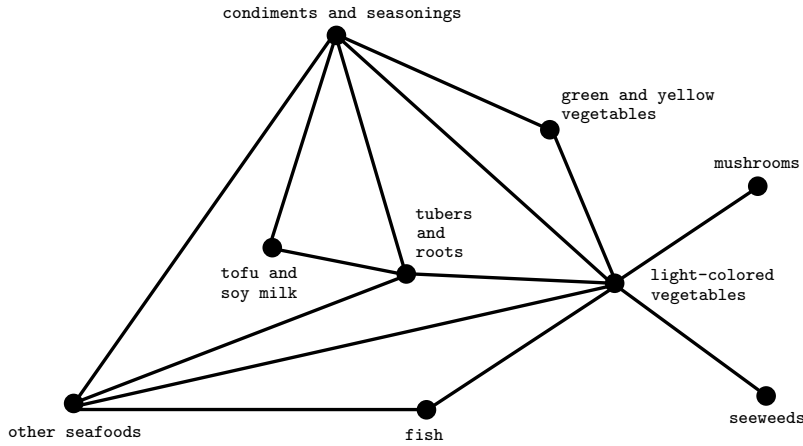


Fig. 2 Main dietary pattern for men from Hoang et al. (2020).

This graph was obtained in Hoang et al. (2020) by applying graphical lasso (Friedman et al., 2008) to a sample of 3 769 subjects, and the estimates of the nonzero partial correlations can be found in Table 4. This sample is part of a larger dataset from a cross-sectional study carried out in South Korea between 2007 and 2019.

In the analysis of dietary patterns it is of interest to identify food groups that play a central role in the eating behaviour (Iqbal et al., 2016; Schwedhelm et al., 2018). For the concentration graph model of Figure 2 we computed the centrality values of the vertices according to the three criteria given above. More specifically, because betweenness centralities scale with the number of pairs of vertices, it is common practice to apply the following normalization,

$$\tilde{B}^x(\cdot) = \frac{B^x(\cdot) - B_{\min}^x}{B_{\max}^x - B_{\min}^x}$$

where B_{\min}^x and B_{\max}^x are the minimal and maximal values of $\tilde{B}^x(\cdot)$, respectively. Hence, the normalized vertex centralities are given in Table 3.

It is of interest to compare the generic basic centrality with the two specific weight and inflation centralities. To this aim we look at the correlation coefficient between every pair of measures, which turns out to be always positive, ranging from 0.79 to 0.91. Hence, from this viewpoint the three measures provide similar results. There are however also some differences of interest. Both the basic and the weight centrality identify **light-color vegetables** as the most central vertex, whereas the inflation centrality puts this vertex in second position, behind **condiment and seasoning**. We can, somehow informally, say that **light-color vegetables** is a central vertex because it contributes to the computation of a high proportion of the correlation of other variables, whereas **condiment and seasoning** is a central vertex because of the number of paths it belongs to and the relevant contribution it gives to

Table 3 Normalized vertex betweenness centralities for the dietary pattern of Figure 2. Vertices are ordered by decreasing value of basic betweenness centrality.

Food group (vertex)	normalized vertex betweenness		
	basic \tilde{B}^ϕ	weight \tilde{B}^ω	inflation \tilde{B}^ν
light-colored vegetables	1.000	1.000	0.747
condiments and seasonings	0.799	0.712	1.000
other seafood	0.686	0.356	0.146
tubers and roots	0.675	0.107	0.180
tofu and soy milk	0.265	0.004	0.013
fish	0.242	0.000	0.002
green and yellow vegetables	0.230	0.071	0.132
mushrooms	0.000	0.000	0.000
seaweeds	0.000	0.000	0.000

the weight of such paths. Furthermore, basic betweenness identifies a cluster of 4 vertices with high centrality value whereas both the weight and inflation centralities restrict the set of highly central vertices to two elements, thereby highlighting the relevance of these two vertices in the network.

We turn now to edge betweenness, whose values are given in Table 4. Unlike vertex centrality, there are important differences in this case. Indeed, the most central vertex according to basic centrality is the least central vertex according to the other two types of edge centrality. More generally, basic centrality is negatively correlated with each of the other two centralities. In concentration graph models, an edge is not present in the graph if its partial correlation is equal to zero, and the absolute value of partial correlations are often regarded as a measure of edge relevance. Partial correlations enter in the computation of path weights whereas they play no role in the computation of the basic centrality. More specifically, we note that the correlation between the values of the basic centrality measures and the estimated of partial correlations is equal to -0.11 , and therefore negative. We also note that, for the most central edge according to the basic centrality, that is the edge joining **tubers and roots** with **other seafood**, the associated partial correlation is one of those with smallest value. Hence, from this perspective, the edge basic centrality measure does not seem to properly suit the graphical model framework. As expected, both weight and inflation centrality have a positive correlation with the estimated partial correlations. On the other hand, partial correlation is only one of the determinant of these centrality values and, interestingly, the two most central vertices according to both the weight and the inflation centralities are the edges joining **light colored vegetables** with **mushrooms** and **seaweeds**, respectively, and the removal of any of these edges would make the graph disconnected. The results provided by the weight and the inflation edge centralities are similar, but not identical, and the correlation between the values of these two measures is equal to 0.7 . When comparing the three edge centralities it is interesting to notice that the five most central edges according to the inflation centrality have all one of the endpoints equal to **light colored vegetables**.

Table 4 Estimated partial correlations and normalized edge betweenness centralities for the dietary pattern of Figure 2. Edges are ordered by decreasing value of basic betweenness centrality.

Edge		$\hat{\rho}_{uv rest}$	normalized edge betweenness		
			basic \tilde{B}^ϕ	weight \tilde{B}^ω	inflation \tilde{B}^ν
tubers and roots	other seafood	0.010	1.000	0.000	0.000
other seafood	fish	0.020	0.677	0.713	0.162
tofu and soy milk	tubers and roots	0.040	0.646	0.158	0.176
condiments and season.	other seafood	0.090	0.600	0.823	0.360
condiments and season.	green and yellow veg.	0.140	0.590	0.589	0.449
condiments and season.	tofu and soy milk	0.040	0.558	0.681	0.272
condiments and season.	tubers and roots	0.100	0.330	0.465	0.526
light-colored vegetables	fish	0.001	0.312	0.109	0.387
light-colored vegetables	tubers and roots	0.030	0.308	0.314	0.475
light-colored vegetables	green and yellow veg.	0.120	0.292	0.548	0.570
light-colored vegetables	mushrooms	0.030	0.110	1.000	1.000
light-colored vegetables	seaweeds	0.020	0.110	1.000	0.992
light-colored vegetables	other seafood	0.040	0.105	0.501	0.592
light-colored vegetables	condiments and season.	0.100	0.000	0.868	0.766

Indeed, the inflation centrality clearly identifies all the edges starting from **light colored vegetables** as highly central so as to pinpoint the relevance of the hub associated with this vertex. On the other hand, the basic edge centrality ranks the vertices of this hub in its lowest positions thereby regarding this structural component of the network as non-central. This seems to be in contradiction with the basic vertex centrality which identifies **light colored vegetables** as the most central vertex. Finally, the information provided by the weight edge centrality with respect to this hub is more ambiguous giving high centrality value to some edges but low value to others.

We close this section by noticing that, potentially, there are exponentially many paths between two vertices of a graph and therefore, for large graphs, the computation of centrality measures that requires the identification of all paths may be computationally unfeasible. The weight and inflation centrality measures introduced in this section seem to give comparable results; however, inflation centrality has the advantage that it is computationally less demanding because its computation does not involve all the paths between two vertices, but only those involving the vertex of interest.

6 Discussion

In recent years there is a growing interest on how to make use and interpret the properties of networks, such as the identification of relevant edges and paths, the computation of centrality measures, the identification of communities. Of special interest is the investigation of methods especially suited for graphical models where the structure of the graph encodes the independence structure of the variables. The theory developed in this paper goes in this direction. Paths play a central role in undirected graphical models and are the key structures to be used in the identification, for instance, of relevant patterns and of vertex which may be regarded as central. It is therefore important to meaningfully associate weights to the paths of a graph which may then be used in the computation of summary measures, such as betweenness centrality measures, and in the comparison of relevant patterns.

In the examples considered in this paper there seems to be a relationship between weight and path length, in the sense that the shorter the path the larger the path weight. This is due to the role played by partial correlations. As shown in (17), if we start from a path and add one edge to it, then the original weight is updated by multiplying it by two factors: (i) an inflation factor that makes the weight value larger because it is greater than one, (ii) a partial correlation that makes the weight value smaller because it belongs to the interval $(-1, 1)$. In the examples we consider, the partial correlation component of the update has always a stronger effect and thus longer paths tend to have smaller weight. A formal analysis of this behavior is an interesting direction of future research so as to clarify to what extent, in large graphs, one could discard large paths and restrict the attention to smaller ones.

The family of undirected graph models and the family of models for directed acyclic graphs (DAGs) have some elements in common. More specifically, there exists a one-to-one relationship between the family of models for undirected decomposable graphs and the family of models for perfect DAGs. In DAGs the relevance of a path is quantified by the theory of path analysis and a second future research direction involve the comparison of the theory of path weights in models which belong to both families.

Acknowledgements The author would like to thank two Referees for providing feedbacks that improved the quality of the paper. We acknowledge the support of Air Force Office of Scientific Research under award number FA9550-17-1-0039.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Belsley DA, Kuh E, Welsch RE (2005) Regression diagnostics: Identifying influential data and sources of collinearity, vol 571. John Wiley & Sons
- Bringmann LF, Elmer T, Epskamp S, Krause RW, Schoch D, Wichers M, Wigman JT, Snippe E (2019) What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology* 128(8):892–903
- Bröhl T, Lehnertz K (2019) Centrality-based identification of important edges in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29(3):033115
- Chatterjee S, Hadi AS (2012) Regression analysis by example, 5th edn. John Wiley & Sons
- Cox DR, Wermuth N (1996) Multivariate Dependencies: Models, analysis and interpretation. Chapman and Hall, London
- Dablander F, Hinne M (2019) Node centrality measures are a poor substitute for causal inference. *Scientific reports* 9(1):1–13
- Dempster AP (1972) Covariance selection. *Biometrics* 28(1):157–175
- Fox J, Monette G (1992) Generalized collinearity diagnostics. *Journal of the American Statistical Association* 87(417):178–183
- Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* 40(1):35–41
- Freeman LC, Borgatti SP, White DR (1991) Centrality in valued graphs: A measure of betweenness based on network flow. *Social networks* 13(2):141–154
- Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9(3):432–441
- Girvan M, Newman ME (2002) Community structure in social and biological networks. *Proceedings of the national academy of sciences* 99(12):7821–7826

- Hoang T, Lee J, Kim J (2020) Differences in dietary patterns identified by the Gaussian graphical model in Korean adults with and without a self-reported cancer diagnosis. *Journal of the Academy of Nutrition and Dietetics*
- Iqbal K, Buijsse B, Wirth J, Schulze MB, Floegel A, Boeing H (2016) Gaussian graphical models identify networks of dietary intake in a german adult population. *The Journal of Nutrition* 146(3):646–652
- Jones B, West M (2005) Covariance decomposition in undirected Gaussian graphical models. *Biometrika* 92(4):779–786
- Lauritzen SL (1996) *Graphical Models*. Oxford University Press
- Lauritzen SL, Richardson TS (2002) Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3):321–348
- Maathuis M, Drton M, Lauritzen S, Wainwright M (2019) *Handbook of Graphical Models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, URL <https://books.google.it/books?id=4x19DwAAQBAJ>
- Newman MJ (2005) A measure of betweenness centrality based on random walks. *Social Networks* 27(1):39–54, DOI <https://doi.org/10.1016/j.socnet.2004.11.009>, URL <https://www.sciencedirect.com/science/article/pii/S0378873304000681>
- Peeters CFW, Bilgrau AE, van Wieringen WN (2020) rags2ridges: A one-stop-shop for graphical modeling of high-dimensional precision matrices. *arXiv:201005619* 2010.05619
- Rodrigues FA (2019) Network centrality: an introduction. In: *A mathematical modeling approach from nonlinear dynamics to complex systems*, Springer, pp 177–196
- Roverato A, Castelo R (2017) The networked partial correlation and its application to the analysis of genetic interactions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 66(3):647–665
- Roverato A, Castelo R (2018) Differential networking with path weights in gaussian trees. In: Kratochvíl V, Studený M (eds) *Proceedings of the Ninth International Conference on Probabilistic Graphical Models*, PMLR, Prague, Czech Republic, *Proceedings of Machine Learning Research*, vol 72, pp 404–415, URL <http://proceedings.mlr.press/v72/roverato18a.html>
- Roverato A, Castelo R (2020) Path weights in concentration graphs. *Biometrika* 107(3):705–722, DOI 10.1093/biomet/asaa010, URL <https://doi.org/10.1093/biomet/asaa010>, <https://academic.oup.com/biomet/article-pdf/107/3/705/33658394/asaa010.pdf>
- Schwedhelm C, Knüppel S, Schwingshackl L, Boeing H, Iqbal K (2018) Meal and habitual dietary networks identified through semiparametric gaussian copula graphical models in a german adult population. *PloS one* 13(8):e0202936
- Whittaker J (1990) *Graphical Models in Applied Multivariate Analysis*. John Wiley & Sons, Chichester
- Wright S (1921) Correlation and causation. *J Agric Res* 20(7):557–585