



Model-based clustering for random hypergraphs

Tin Lok James Ng¹ · Thomas Brendan Murphy²

Received: 7 July 2020 / Revised: 14 June 2021 / Accepted: 20 June 2021 / Published online: 28 June 2021
© The Author(s) 2021

Abstract

A probabilistic model for random hypergraphs is introduced to represent unary, binary and higher order interactions among objects in real-world problems. This model is an extension of the latent class analysis model that introduces two clustering structures for hyperedges and captures variation in the size of hyperedges. An expectation maximization algorithm with minorization maximization steps is developed to perform parameter estimation. Model selection using Bayesian Information Criterion is proposed. The model is applied to simulated data and two real-world data sets where interesting results are obtained.

Keywords Clustering · Hypergraph · Latent class analysis · Minorization maximization

Mathematics Subject Classification 62H99 · 62P25

1 Introduction

A large number of random graph models have been proposed (Nowicki and Snijders 2001; Hoff et al. 2002; Handcock et al. 2007; Latouche et al. 2011) to describe complex interactions among objects of interest. Pairwise relationships among objects can be naturally represented as a graph, in which the objects are represented by the vertices, and two vertices are joined by an edge if certain relationship exists between them. While graphs are capable of representing pairwise interactions between objects, they are inadequate to represent unary and higher order interactions that are typically observed in many real-world problems. Examples of data with unary and higher-order interactions include co-authorship on academic papers, co-appearance in movie scenes, and songs performed in a concert.

✉ Tin Lok James Ng
jamesng1105@gmail.com

¹ School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland

² School of Mathematics and Statistics, University College Dublin, Dublin, Ireland

For example, the study of coauthorship networks of scientists have attracted significant interest in both natural and social sciences (Newman 2001a, b, 2004; Moody 2004; Azondekon et al. 2018). Such networks are typically constructed by connecting two scientists if they have coauthored one or more papers together. However, as we will illustrate below, such representation inevitably results in loss of information while a hypergraph representation naturally preserves all information. A hypergraph is a generalization of a graph in which hyperedges are arbitrary sets of vertices, and can contain any number of vertices. As a result, hypergraphs are capable of representing relationships of any order.

We consider a simple example of a coauthorship network with 7 authors and 4 papers, in order to illustrate the benefits of hypergraph modelling. A hypergraph representation of the network is given in Fig. 1, where the vertices v_1, v_2, \dots, v_7 represent the authors while the hyperedges e_1, \dots, e_4 represent the papers. For example, the paper e_1 is written by four authors v_1, v_2, v_3 and v_4 , the paper e_2 is written by two authors v_2 and v_3 , the paper e_3 has v_3, v_5 and v_6 as authors and the paper e_4 has a single author v_4 .

On the other hand, a graph representation of this coauthorship network with edges between any two authors who have coauthored at least one paper results in the edge set $\{(v_1, v_2), (v_1, v_3), (v_1, v_4), (v_2, v_3), (v_2, v_4), (v_3, v_4), (v_3, v_5), (v_3, v_6), (v_5, v_6)\}$. It is evident that much information is lost with this representation. In particular, this representation removes information about the number of authors that co-authored a paper. For example, one can only deduce from this edge set that v_3 has co-authored with v_1 and v_2 while unable to conclude that the co-authorship was for the same paper. Furthermore, the hyperedge e_4 which contains a singleton v_4 is left out in the graph representation.

A number of random hypergraph models have been studied in probability and combinatorics literature, where theoretical properties are investigated (Karoński and Łuczak 2002; Goldschmidt 2005; de Panafieu 2015; Dyer et al. 2015; Poole 2015). A novel parametrization of distributions on hypergraphs based on the geometry of points is proposed in Lunagómez et al. (2017) which is used to infer Markov structure for multivariate distributions. On the other hand, statistical modeling of random hypergraph data is less developed. Stasi et al. (2014) introduced the hypergraph beta model with three variants, which is a natural extension of the beta model for random graphs (Holland and Leinhardt 1981). In their model, the probability of a hyperedge e appearing in the hypergraph is parameterized by a vector $\beta \in \mathbf{R}^N$, which represents the “attractiveness” of each vertex. However, their model does not capture clustering among objects, which is a typical real world phenomenon. In addition, the assumption of an upper bound on the size of hyperedges violates the structure of many real world data sets.

One may equivalently represent a hypergraph using a bipartite network (also called two-mode network and affiliation network). Two-mode networks consist of two different kinds of vertices and edges can only be observed between the two types of vertices, but not between vertices of the same type. A hypergraph can be represented as a two-mode network by considering the hyperedges as a second type of vertices. For example, an equivalent bipartite representation of the hypergraph shown in Fig. 1

is provided in Fig. 2 where the hyperedges $\{e_1, \dots, e_4\}$ are now replaced by the four green vertices.

Two-mode networks have been studied in various disciplines including computer science (Perugini et al. 2004), social sciences (Faust et al. 2002; Koskinen and Edling 2012; Friel et al. 2016) and physics (Lind et al. 2005). A number of approaches have been proposed to analyze and model two-mode network data (Borgatti and Everett 1997; Doreian and Batagelj 2004; Latapy et al. 2008; Wang et al. 2009; Snijders et al. 2013; Aitkin et al. 2014). In particular, models originally developed for binary networks were extended for two-mode networks.

Doreian and Batagelj (2004) develops a blockmodeling approach of two-mode network data which aims to simultaneously partition the two types of vertices into blocks. Skvoretz and Faust (1999) proposes the exponential random graph models (ERGMs) for two-mode networks, which models the logit of the probability of an actor belonging to an event as a function of actor and event specific effects and other graph statistics. A clustering algorithm for two-mode networks is developed in Field et al. (2006) based on the modelling framework in Skvoretz and Faust (1999). Several extensions to the ERGMs for bipartite networks are proposed by (eg. Wang et al. 2009, 2013). Snijders et al. (2013) proposes a methodology for studying the co-evolution of two-mode and one-mode networks. A network autocorrelation model for two-mode networks is introduced in Fujimoto et al. (2011). Aitkin et al. (2014) evaluates the identification of clustering structure in bipartite networks through latent class analysis and introduces a new Bayesian method for choosing the number of latent classes.

Representing network observations using two-mode networks has the benefits of modelling vertices of both types jointly. However, in analyzing a two-mode network, one type of vertices may attract most interest. For example, in co-authorship networks, the main interest may lie in the collaborations rather than in co-authored papers. When modeling the co-appearance of characters in the scenes of a movie, one is typically interested in co-appearance of the characters rather than the movie scenes. In such scenarios, a hypergraph representation is most natural by converting one type of vertex into hyperedge.

A related and popular research problem is hypergraph partitioning (Zhou et al. 2007; Leordeanu and Sminchisescu 2012; Purkait et al. 2017). Hypergraph partitioning aims to partition vertices in a hypergraph into clusters based on their higher order interactions, and is an important research problem in computer vision (Agarwal et al. 2006; Li et al. 2013), recommender systems (Bu et al. 2010) and other fields. In contrast, we propose a random hypergraph model which captures the clustering structure of the hyperedges. Since hyperedges are simply arbitrary sets of vertices, interpretable structure within the vertices can also be inferred from the clustering structure of the hyperedges. By adopting a probabilistic approach to hypergraph modeling, the proposed model is capable of quantifying the uncertainties in the clustering of hyperedges.

In this paper, we propose the Extended Latent Class Analysis (ELCA) model for random hypergraphs, which is a natural extension of the Latent Class Analysis (LCA) model (Lazarsfeld and Henry 1968; Goodman 1974; Celeux and Govaert 1991) and includes the LCA model as a special case. The ELCA can alternatively be interpreted as a constrained case of the LCA and it achieves significant reduction in model complexity. Furthermore, the model directly captures the variation in sizes of hyperedges

Fig. 1 A hypergraph representation of a coauthorship network

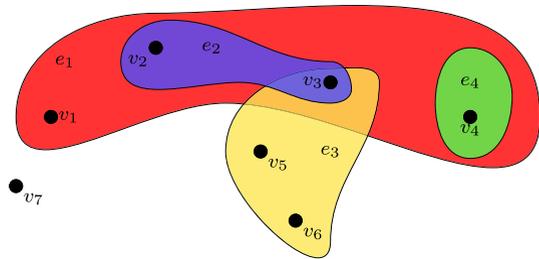
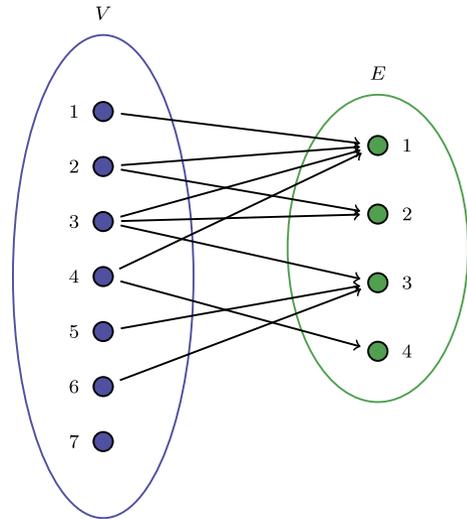


Fig. 2 Bipartite graph representation of the hypergraph in Fig. 1



which are typically observed in applications. For example, the number of authors per scientific publication varies widely across different disciplines. We develop an EM (Expectation Maximization) algorithm with MM (Minorization Maximization) steps to perform parameter estimation. To determine the number of latent classes, we employ the Bayesian Information Criterion (BIC). The model is applied to simulated data, and two applications: Star Wars movie scenes and Reuters news articles.

2 Model and motivation

2.1 Hypergraph

A hypergraph is represented by a pair $H = (V, E)$, where $V = \{v_1, v_2, \dots, v_N\}$ is the set of N vertices and $E = \{e_1, e_2, \dots, e_M\}$ is the set of M hyperedges. A hyperedge e is a subset of V , and we allow repetitions in the hyperedge set E . Thus, the hypergraph H can alternatively be represented with a $N \times M$ matrix $\mathbf{X} = (x_{ij})$ where $x_{ij} = 1$ if vertex v_i appears in hyperedge e_j and $x_{ij} = 0$ otherwise.

2.2 Latent class analysis model for random hypergraphs

The binary latent class analysis (LCA) model (Lazarsfeld and Henry 1968; Goodman 1974) is a commonly used mixture model for high dimensional binary data. It assumes that each observation is a member of one and only one of the G latent classes, and conditional on the latent class membership, the manifest variables are mutually independent of each other. The LCA model appears to be a natural candidate to model random hypergraphs where hyperedges are partitioned into G latent classes, and the probability that a hyperedge $e \in E$ contains a vertex $v \in V$ depends only on its latent class assignment.

Let $\mathbf{X} = (x_{ij})$ be the matrix representation of the hypergraph H where $x_{ij} = 1$ if vertex v_i is in hyperedge e_j and $x_{ij} = 0$ otherwise. Let $\pi = (\pi_1, \dots, \pi_G)$ be the *a priori* latent class membership probabilities, where π_g is the probability that a hyperedge belongs to latent class g . We define the $N \times G$ matrix \mathbf{P} , where p_{ig} is the probability that vertex v_i is contained in a hyperedge e with latent class membership g . The probability of observed hyperedge e_j , which is represented by (x_{1j}, \dots, x_{Nj}) , is thus

$$\sum_{g=1}^G \pi_g \prod_{i=1}^N p_{ig}^{x_{ij}} (1 - p_{ig})^{1-x_{ij}}.$$

Thus, the likelihood function of \mathbf{P} and π can be written as

$$L(\mathbf{X}; \mathbf{P}, \pi) = \prod_{j=1}^M \left[\sum_{g=1}^G \pi_g \prod_{i=1}^N p_{ig}^{x_{ij}} (1 - p_{ig})^{1-x_{ij}} \right].$$

Let $\mathbf{Z}^{(1)}$ be a $M \times G$ latent class membership matrix, where $z_{jg}^{(1)} = 1$ if hyperedge e_j has latent class label g and $z_{jg}^{(1)} = 0$ otherwise. The complete-data likelihood of \mathbf{P} and π can be expressed as (1).

$$L(\mathbf{X}, \mathbf{Z}^{(1)}; \mathbf{P}, \pi) = \prod_{j=1}^M \prod_{g=1}^G \left[\pi_g \prod_{i=1}^N p_{ig}^{x_{ij}} (1 - p_{ig})^{1-x_{ij}} \right]^{z_{jg}^{(1)}}. \tag{1}$$

In comparison to the hypergraph beta models introduced in Stasi et al. (2014), the LCA model is capable of capturing the clustering and heterogeneity of hyperedges. For example, academic papers can be naturally labelled according to subject areas and conditional on a paper being labelled mathematics, one would expect that the probability a mathematician co-authored the paper is higher than a biologist. The LCA model does not assume an upper bound on the size of hyperedges and can model hyperedges of any size. Furthermore, an expectation maximization algorithm (Dempster et al. 1977) can be easily derived to perform parameter estimation.

2.3 Extended latent class analysis for random hypergraphs

While the LCA model captures the clustering and heterogeneity of hyperedges in real world data sets, a large number of latent classes are typically required to achieve a good fit of the data. As a result, the number of parameters grows quickly with a moderate or large number of nodes. The complexity of the LCA model can be substantially reduced if we assume that some of the latent class conditional probabilities $(p_{ig})_{i=1}^N$ tend to be proportional to each other for different values of g . While assuming proportionality of latent class conditional probabilities may appear rather restrictive, it is a reasonable assumption in many hypergraph applications. We develop the Extended Latent Class Analysis (ELCA) model which builds on the proportionality assumption on the conditional probabilities.

Let $a = (a_1, \dots, a_K)$ with $0 \leq a_k \leq 1$ be a K dimensional vector, the ELCA model assumes that the latent class conditional probabilities are of the form $(\phi_{ig}a_k)_{i=1}^N$ for $g = 1, \dots, G$ and $k = 1, \dots, K$. In the context of hypergraph applications, the a_k parameters capture the variations in the size (number of vertices) of the hyperedges whereas the ϕ_{ig} values capture the probability that a node is contained in a hyperedge. The ELCA model can be considered as having two types of clustering structure, with the primary clustering structure defined by ϕ_{ig} parameters and an additional clustering structure captured by a_k parameters. We note that the ELCA reduces to the standard LCA when $K = 1$.

Let $\tau = (\tau_1, \dots, \tau_K)$ be the clustering assignment probabilities corresponding to the additional structure, the ELCA model assumes that these two clustering structure are *a priori* independent. Thus, the probability that a hyperedge has primary cluster label g and additional cluster label k is $\pi_g \tau_k$, and the probability that the vertex v_i is contained in a hyperedge from the clusters pair (g, k) is $a_k \phi_{ig}$, and the probability that the vertex v_i is contained in a hyperedge from the primary cluster g is $\phi_{ig} \sum_{k=1}^K a_k \tau_k$.

Under the ELCA model with G primary clusters and K additional clusters, the probability of observing a hyperedge (x_{1j}, \dots, x_{Nj}) is given by

$$\sum_{g=1}^G \sum_{k=1}^K \pi_g \tau_k \prod_{i=1}^N (a_k \phi_{ig})^{x_{ij}} (1 - a_k \phi_{ig})^{1-x_{ij}}.$$

Let $\theta = (\pi, \tau, \phi, a)$ denote the model parameters, the likelihood function can be written as

$$L(\mathbf{X}; \theta) = \prod_{j=1}^M \left\{ \sum_{g=1}^G \sum_{k=1}^K \pi_g \tau_k \prod_{i=1}^N (a_k \phi_{ig})^{x_{ij}} (1 - a_k \phi_{ig})^{1-x_{ij}} \right\}.$$

The ELCA model is not identifiable if the parameters $(a_k)_{k=1}^K$ are not constrained. To see this, if $0 < a_k < 1$ for all k , then the likelihood function is invariant under the transformation $(a_k)_{k=1}^K \rightarrow (Ca_k)_{k=1}^K$ and $(\phi_{ig})_{i=1, g=1}^{i=N, g=G} \rightarrow (C^{-1} \phi_{ig})_{i=1, g=1}^{i=N, g=G}$, where C is some positive constant such that $\max_k \{Ca_k\} \leq 1$. Thus, to ensure the identifiability of the model, $(a_k)_{k=1}^K$ are ranked by increasing order with $a_K = 1$.

We define the $M \times K$ additional cluster membership matrix $\mathbf{Z}^{(2)} = (z_{jk}^{(2)})$, where $z_{jk}^{(2)} = 1$ if hyperedge e_j has additional cluster label k and $z_{jk}^{(2)} = 0$ otherwise. The complete data likelihood function of \mathbf{X} , $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ is given as

$$L(\mathbf{X}, \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}; \theta) = \prod_{j=1}^M \prod_{g=1}^G \prod_{k=1}^K \left[\pi_g \tau_k \prod_{i=1}^N (a_k \phi_{ig})^{x_{ij}} (1 - a_k \phi_{ig})^{1-x_{ij}} \right]^{z_{jg}^{(1)} z_{jk}^{(2)}}. \tag{2}$$

We note that any ELCA with G primary clusters and K additional clusters can be equivalently represented as a standard LCA with $G \times K$ clusters. Under the standard LCA representation of the ELCA model, the $G \times K$ vectors of latent class conditional probabilities $\{(p_{ig})_{i=1}^N\}_{g=1}^{G \times K}$ can be partitioned into G sets of equal size K , and $(p_{ig})_{i=1}^N$ are proportional to each other within each set with the constants of proportionality determined by $(a_k)_{k=1}^K$. Consider the ELCA with 2 primary clusters and 2 additional clusters, which is a special case of the 4-cluster LCA model. The probabilities that vertex v_i is contained in a hyperedge from the cluster pair $(1, 1)$, $(1, 2)$, $(2, 1)$, $(2, 2)$ are given by ϕ_{i1} , ϕ_{i2} , $a_1 \phi_{i1}$, $a_1 \phi_{i2}$.

It is easy to see that under the proportionality assumption, the ELCA model achieves significant reduction in the number of parameters. For the ELCA model with G primary clusters and K additional clusters, the number of parameters is given by $GN + 2(K - 1) + (G - 1)$ whereas the number of parameters for the LCA with $G \times K$ clusters is $GKN + (GK - 1)$.

2.4 Theoretical properties

We analyze the distribution of the size of a random hyperedge under the proposed ELCA model. Proposition 1 below shows that the size of the hyperedges simulated from the ELCA model tend to have larger variance than those simulated from the LCA model.

Proposition 1 *Suppose we are given the LCA model with parameters $\{\pi, p\}$ and the ELCA model with parameters $\{\pi, \tau, a, \phi\}$ and N vertices. Suppose the condition $p_{ig} = \phi_{ig} \sum_{k=1}^K a_k \tau_k$ holds for $i = 1, \dots, N$ and $g = 1, \dots, G$. This condition ensures that the latent class conditional probabilities of the primary clustering structure are the same for both models.*

Let A denote the cardinality $|e_1|$ of a random hyperedge e_1 generated under the LCA model. Similarly, let B denote the cardinality $|e_2|$ of a random hyperedge e_2 generated under the ELCA model. We have the following results:

$$E(A) = E(B)$$

$$Var(A) \leq Var(B).$$

Proof The proof is straightforward and is given in the Appendix. □

We now let $f_N(y)$ be the probability mass of the size of a random hyperedge simulated from a G cluster LCA model. Similarly, we let $h_N(y)$ be the probability

mass of the size of a random hyperedge simulated from the ELCA model with G clusters and K additional clusters. The following result can be derived.

Proposition 2 1. Under the specifications of a LCA model with parameters $\pi = (\pi_1, \dots, \pi_G)$ and $\{p_{ig}\}_{i=1, \dots, N, g=1, \dots, G}$, and suppose the following conditions hold for $g = 1, \dots, G$,

$$\lambda_N^{(g)} = \sum_{i=1}^N p_{ig} \rightarrow \lambda^{(g)} > 0$$

$$\sum_{i=1}^N p_{ig}^2 \rightarrow 0$$

as $N \rightarrow \infty$. We have

$$f_N(y) \rightarrow \sum_{g=1}^G \pi_g \frac{e^{-\lambda^{(g)}} (\lambda^{(g)})^y}{y!}.$$

That is, the distribution of the size of a random hyperedge converges to a mixture of Poisson distributions with G components.

2. Under the specification of a ELCA model with parameters $\pi = (\pi_1, \dots, \pi_G)$, $\tau = (\tau_1, \dots, \tau_K)$, $a = (a_1, \dots, a_K)$, and $\{\phi_{ig}\}_{i=1, \dots, N, g=1, \dots, G}$, and further suppose the following conditions hold for $g = 1, \dots, G$, and $k = 1, \dots, K$,

$$\lambda_N^{(g,k)} = \sum_{i=1}^N \phi_{ig} a_k \rightarrow \lambda^{(g,k)} > 0$$

$$\sum_{i=1}^N \phi_{ig}^2 a_k^2 \rightarrow 0$$

as $N \rightarrow \infty$. We have

$$h_N(y) \rightarrow \sum_{g=1}^G \sum_{k=1}^K \pi_g \tau_k \frac{e^{-\lambda^{(g,k)}} (\lambda^{(g,k)})^y}{y!}.$$

That is, the distribution of the size of a random hyperedge converges to a mixture of Poisson distributions with $G \times K$ components.

Proof Conditional on the event that a random hyperedge is generated from cluster g , (Wang 1993, Theorem 3) implies that

$$f_N(y) \rightarrow \frac{e^{-\lambda^{(g)}} (\lambda^{(g)})^y}{y!}.$$

Part 1 result follows by marginalizing over the G clusters. The second part of the proposition can be proved similarly. \square

Proposition 2 implies that under mild conditions, the distribution of the size of hyperedges converges to a mixture of Poisson distributions with $G \times K$ mixture components as the number of vertices increases. We note that the mixture components of the limiting mixture of Poisson distribution are subject to the same proportionality condition. Nevertheless, larger variations in the size of hyperedges tend to be obtained under the ELCA compared to those obtained under the standard LCA.

3 Estimation and model selection

3.1 EM algorithm

We estimate the parameters $\theta = (\pi, \tau, \phi, a)$ of the ELCA model using an EM algorithm (Dempster et al. 1977) which is a popular method in fitting mixture models. The E-step of the EM algorithm involves computing the expected value of the complete data log-likelihood (2) with respect to the distribution of the unobserved $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ given the current estimates. The M-step involves maximizing the expected complete data log-likelihood.

Taking logarithm of the complete data likelihood in (2), we obtain the complete data log-likelihood function below.

$$\log L(\mathbf{X}, \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}; \theta) = \sum_{j=1}^M \sum_{g=1}^G \sum_{k=1}^K z_{jg}^{(1)} z_{jk}^{(2)} \left[\log \pi_g + \log \tau_k + \sum_{i=1}^N \left\{ x_{ij} \log(a_k) + \log(\phi_{ig}) + (1 - x_{ij}) \log(1 - a_k \phi_{ig}) \right\} \right]. \tag{3}$$

3.1.1 E-step

For the E-step, we need to evaluate the expected complete data log-likelihood, which is the expectation of (3) conditional on data x and current parameter estimates $\theta^{(t)}$. The expected complete data log-likelihood is denoted as $Q(\theta|\theta^{(t)})$ and is defined as

$$Q(\theta|\theta^{(t)}) := E(\log L(\mathbf{X}, \mathbf{Z}^{(1)}, \mathbf{Z}^{(2)}; \theta) | \mathbf{X}, \theta^{(t)}). \tag{4}$$

Because the complete-data log-likelihood is linear in $Z_{jg}^{(1)} Z_{jk}^{(2)}$, we need to evaluate the expectation $\widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} := E(Z_{jg}^{(1)} Z_{jk}^{(2)} | \mathbf{X}, \theta^{(t)})$. We have that

$$\begin{aligned} E(Z_{jg}^{(1)} Z_{jk}^{(2)} | \mathbf{X}, \theta^{(t)}) &= Pr(Z_{jg}^{(1)} = Z_{jk}^{(2)} = 1 | \mathbf{X}, \theta^{(t)}) \\ &= \frac{\pi_g^{(t)} \tau_k^{(t)} \left[\prod_{i=1}^N (a_k \phi_{ig})^{x_{ij}} (1 - a_k \phi_{ig})^{1-x_{ij}} \right]}{\sum_{g=1}^G \sum_{k=1}^K \pi_g^{(t)} \tau_k^{(t)} \left[\prod_{i=1}^N (a_k \phi_{ig})^{x_{ij}} (1 - a_k \phi_{ig})^{1-x_{ij}} \right]}. \end{aligned} \tag{5}$$

In particular, the E-step has a computational complexity of $\mathcal{O}(N)$ for each pair (g, k) , and an overall complexity of $\mathcal{O}(NGK)$.

3.1.2 M-step

While the E-step of the EM algorithm is straightforward, the M-step involves complicated maximization. For the M-step, we need to maximize $Q(\theta|\theta^{(t)})$ with respect to the model parameters $\{\phi_{ig}\}$, $\{a_k\}$, $\{\pi_g\}$ and $\{\tau_k\}$. Thus, we use the ECM algorithm (Meng and Rubin 1993) which replaces the complex M-step by a series of simpler conditional maximizations. The conditional maximizations with respect to the parameters ϕ and a do not have closed form solutions. We utilize the MM algorithm (Lange et al. 2000; Hunter and Lange 2004) which works by lower bounding the objective function by a minorizing function and then maximizing the minorizing function. Since the M-step involves a series of conditional maximization, the Q function is guaranteed to increase (Meng and Rubin 1993, Theorem 1).

Maximize w.r.t. ϕ_{ig}

For fixed i and g , the objective function retaining terms involving ϕ_{ig} can be written as

$$Q = \sum_{j=1}^M \sum_{k=1}^K \widehat{Z}_{jg}^{(1)} \widehat{Z}_{jk}^{(2)} \left(x_{ij} \log(\phi_{ig}) + (1 - x_{ij}) \log(1 - a_k \phi_{ig}) \right). \tag{6}$$

An analytic expression for $\arg \max_{\phi_{ig}} \{Q\}$ does not exist due to the $\log(1 - a_k \phi_{ig})$ term and thus we apply the MM (Minorization Maximization) algorithm (Hunter and Lange 2004). We first apply a quadratic lower bound on the concave function $\log(1 - a_k \phi_{ig})$ for $k < K$:

$$\begin{aligned} \log(1 - a_k \phi_{ig}) &\geq \log(1 - a_k \phi_{ig}^{(t)}) + \left(\frac{-a_k}{1 - a_k \phi_{ig}^{(t)}} \right) (\phi_{ig} - \phi_{ig}^{(t)}) \\ &\quad + \frac{1}{2} \left(\frac{-a_k^2}{(1 - a_k)^2} \right) (\phi_{ig} - \phi_{ig}^{(t)})^2. \end{aligned}$$

Hence, the objective function in (6) up to an additive constant can be minorized by Q_{lower} :

$$\begin{aligned} Q_{lower} &= \sum_{j=1}^M \sum_{k=1}^K \widehat{Z}_{jg}^{(1)} \widehat{Z}_{jk}^{(2)} x_{ij} \log(\phi_{ig}) + \sum_{j=1}^M \sum_{k=1}^{K-1} \widehat{Z}_{jg}^{(1)} \widehat{Z}_{jk}^{(2)} (1 - x_{ij}) \\ &\quad \left(\left(\frac{-a_k}{1 - a_k \phi_{ig}^{(t)}} \right) \phi_{ig} + \frac{1}{2} \left(\frac{-a_k^2}{(1 - a_k)^2} \right) (\phi_{ig} - \phi_{ig}^{(t)})^2 \right) \\ &\quad + \sum_{j=1}^M \widehat{Z}_{jg}^{(1)} \widehat{Z}_{jK}^{(2)} (1 - x_{ij}) \log(1 - \phi_{ig}). \end{aligned} \tag{7}$$

To simplify the expression above, we define the quantities below:

$$\begin{aligned}
 A_1 &= \sum_{j=1}^M \sum_{k=1}^K \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} x_{ij} \\
 A_2 &= \sum_{j=1}^M \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} (1 - x_{ij}) \\
 B_1 &= \sum_{j=1}^M \sum_{k=1}^{K-1} \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} (1 - x_{ij}) \frac{-a_k}{1 - a_k \phi_{ig}^{(t)}} \\
 B_2 &= \sum_{j=1}^M \sum_{k=1}^{K-1} \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} (1 - x_{ij}) \frac{1}{2} \frac{-a_k^2}{(1 - a_k)^2}.
 \end{aligned}$$

Now, the lower bound (7) can be written as below.

$$Q_{lower} = A_1 \log(\phi_{ig}) + A_2 \log(1 - \phi_{ig}) + B_1 \phi_{ig} + B_2 (\phi_{ig} - \phi_{ig}^{(t)})^2.$$

Taking derivative with respect to ϕ_{ig} , we have

$$\frac{A_1}{\phi_{ig}} - \frac{A_2}{1 - \phi_{ig}} + B_1 + 2B_2 \phi_{ig} - 2B_2 \phi_{ig}^{(t)} = 0.$$

Let $C = B_1 - 2B_2 \phi_{ig}^{(t)}$, we have

$$\phi_{ig}^3 - \frac{2B_2 - C}{2B_2} \phi_{ig}^2 - \frac{C - A_1 - A_2}{2B_2} \phi_{ig} - \frac{A_1}{2B_2} = 0. \tag{8}$$

Solving the cubic equation above results in the update for ϕ_{ig} .

Maximize w.r.t. a_k

For a fixed k , the objective function (3) retaining terms involving a_k can be expressed as

$$Q = \sum_{j=1}^M \sum_{g=1}^G \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} \left(\sum_{i=1}^N x_{ij} \log(a_k) + (1 - x_{ij}) \log(1 - a_k \phi_{ig}) \right). \tag{9}$$

Since an analytic expression for $\arg \max_{a_k} \{Q\}$ does not exist due to the $\log(1 - a_k \phi_{ig})$ term, we apply the MM (Minorization Maximization) algorithm. We first apply a quadratic lower bound on the concave function

$$\begin{aligned} \log(1 - a_k \phi_{ig}) &\geq \log(1 - a_k^{(t)} \phi_{ig}) + \left(\frac{-\phi_{ig}}{1 - a_k^{(t)} \phi_{ig}} \right) (a_k - a_k^{(t)}) \\ &\quad + \frac{1}{2} \left(\frac{-\phi_{ig}^2}{(1 - \phi_{ig})^2} \right) (a_k - a_k^{(t)})^2. \end{aligned}$$

Hence, (9) up to an additive constant can be minorized by the function:

$$\begin{aligned} Q_{lower} &= \left(\sum_{j=1}^M \sum_{g=1}^G \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} \sum_{i=1}^N x_{ij} \right) \log(a_k) + \sum_{j=1}^M \sum_{g=1}^G \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} \sum_{i=1}^N (1 - x_{ij}) \\ &\quad \left(\frac{-\phi_{ig}}{1 - a_k^{(t)} \phi_{ig}} a_k + \frac{1}{2} \left(\frac{-\phi_{ig}^2}{(1 - \phi_{ig})^2} \right) (a_k - a_k^{(t)})^2 \right). \end{aligned} \tag{10}$$

To simplify the expression above, we define the following quantities:

$$\begin{aligned} A &= \sum_{j=1}^M \sum_{g=1}^G \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} \sum_{i=1}^N x_{ij} \\ B &= \sum_{j=1}^M \sum_{g=1}^G \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} \sum_{i=1}^N (1 - x_{ij}) \left(\frac{-\phi_{ig}}{1 - a_k^{(t)} \phi_{ig}} \right) \\ C &= \sum_{j=1}^M \sum_{g=1}^G \widehat{Z_{jg}^{(1)} Z_{jk}^{(2)}} \sum_{i=1}^N (1 - x_{ij}) \frac{1}{2} \left(\frac{-\phi_{ig}^2}{(1 - \phi_{ig})^2} \right). \end{aligned}$$

Taking derivative of (9) with respect to a_k , we have

$$\frac{\partial Q_{lower}}{\partial a_k} = \frac{A}{a_k} + B + 2C(a_k - a_k^{(t)}) = 0.$$

Let $D = (\frac{B}{2C} - a_k^{(t)})$, $E = -\frac{A}{2C}$, we have

$$\hat{a}_k = \left(E + \frac{D^2}{4} \right)^{1/2} - \frac{D}{2}. \tag{11}$$

Maximize w.r.t. π_g and τ_k We apply the method of Lagrange multipliers to derive the updates for π_g and τ_k . The objective function for $(\pi_h)_{h=1}^G$ is given by

$$\sum_{j=1}^M \sum_{h=1}^G \sum_{k=1}^K \widehat{z_{jh}^{(1)} z_{jk}^{(2)}} \log \pi_h + \lambda \left(1 - \sum_{h=1}^G \pi_h \right), \tag{12}$$

where λ is the Lagrange multiplier. Differentiating w.r.t. π_g and setting to 0 gives

$$\frac{\sum_{j=1}^M \sum_{k=1}^K \widehat{z_{jg}^{(1)} z_{jk}^{(2)}}}{\pi_g} - \lambda = 0.$$

Therefore, the update for π_g is given by

$$\hat{\pi}_g \propto \sum_{j=1}^M \sum_{k=1}^K \widehat{z_{jg}^{(1)} z_{jk}^{(2)}}. \tag{13}$$

The update for τ_k can be derived analogously and is given below:

$$\hat{\tau}_k \propto \sum_{j=1}^M \sum_{g=1}^G \widehat{z_{jg}^{(1)} z_{jk}^{(2)}}. \tag{14}$$

The EM algorithm is summarized in Algorithm 1, where line 4 corresponds to the expectation step and line 5 - 18 are the conditional maximization steps. In particular, we note that the computational complexity for maximizing ϕ_{ig} and a_k are given by $\mathcal{O}(N_{iter}MK)$ and $\mathcal{O}(N_{iter}MGN)$, respectively, where N_{iter} is the number of iterations required for the MM algorithm.

Algorithm 1 EM Algorithm

Input: \mathbf{x}, G, K, tol

Output: $\hat{\phi}, \hat{a}, \hat{\pi}, \hat{\tau}, \hat{\mathbf{z}}^{(1)}, \hat{\mathbf{z}}^{(2)}$

- 1: *conv* = *False*
 - 2: Random initialization of ϕ, a, π, τ
 - 3: **while** *conv* = *False* **do**
 - 4: Do the E-step according to (5)
 - 5: **for** $i = 1, \dots, N$ **do**
 - 6: **for** $g = 1, \dots, G$ **do**
 - 7: Update ϕ_{ig} according to (8)
 - 8: **end for**
 - 9: **end for**
 - 10: **for** $k = 1, \dots, K - 1$ **do**
 - 11: Update a_k according to (11)
 - 12: **end for**
 - 13: **for** $g = 1, \dots, G$ **do**
 - 14: Update π_g according to (13)
 - 15: **end for**
 - 16: **for** $k = 1, \dots, K$ **do**
 - 17: Update τ_k according to (14)
 - 18: **end for**
 - 19: Evaluate Change in log-likelihood Δ_{loglik} resulting from parameter updates
 - 20: **if** $\Delta_{loglik} < tol$ **then**
 - 21: *conv* = *True*
 - 22: **end if**
 - 23: **end while**
-

3.2 Model selection

We use the Bayesian Information Criterion (BIC) (Schwarz 1978) to determine the optimal number of primary and additional clusters for the ELCA model. For the ELCA model, the BIC takes the following form:

$$-2 \log L + (GN + 2(K - 1) + (G - 1)) \log M$$

where $\log L$ is the log-likelihood evaluated at the estimated parameters, and $GN + 2(K - 1) + (G - 1)$ is the number of parameters in the model. The model with the lowest BIC value is selected. The accuracy of the BIC as a model selection criterion requires M to be relatively large compared to N . For the standard latent class models, existing literature suggests that the BIC is a good indicator of the true number of classes (Collins et al. 1993) and extensive simulation studies were performed in Nylund et al. (2007) to validate this claim. The performance of BIC as a model selection criterion for the ELCA model is assessed using simulation studies in Sect. 4.

4 Simulation studies

We conduct simulation studies to examine the performance of the proposed EM algorithm for the ELCA model and the behavior of BIC as a model selection criterion. The results presented in Tables 1 and 2 are concerned with assessing the convergence behavior of the proposed EM algorithm with various latent class assignment probabilities for primary and additional clusters. Hyperedges are simulated from the ELCA model with two primary clusters and two additional clusters in Table 1 and from the ELCA model with three primary clusters and two additional clusters in Table 2. The specific model parameters used in the simulation are given in the Appendix.

For the model parameters ϕ , a , π and τ of the ELCA model, the ℓ_2 distances between the true parameters and the estimated ones are presented in Tables 1 and 2. The misclassification rates for both the primary and additional clusters are also presented. We observe that the estimated parameters converge to the true values as the number of hyperedges increases. It is worth noting that the convergence tends to be faster in the case of two primary clusters compared to three primary clusters.

We examine the performance of BIC in choosing the optimal number of primary and additional clusters. The values in Tables 3 and 4 are computed by comparing the BIC across a range of models, then identifying where the lowest values occurred across these models considered. The model parameters which generate the hyperedges are given in Appendix. For example, with 10 vertices and 200 hyperedges, the lowest values of BIC occurred at the two primary and two additional cluster model (which is the true model) 67% of the time. Looking across the values in Tables 3 and 4, we notice that the BIC tends to be a less accurate model selection criterion when the number of hyperedges is small but improves significantly as the number of hyperedges M increases.

As a final simulation study, we simulate hyperedges from the LCA models with two and three clusters and note that they are special cases of the ELCA models with

$K = 1$ additional cluster. The simulated data is then fitted with the ELCA models with $K = 2$ and $K = 3$ additional clusters. For various simulation settings, we simulate 100 sets of hyperedges and examine the proportion of times that the true model can be recovered. The true model is considered to be recovered if the estimated parameters satisfy $\max\{\tau_k\} > 1 - \epsilon$ or $\min\{a_k\} > 1 - \epsilon$ for some small positive number ϵ . Simulation results are shown in Table 5 with ϵ is set to 0.01 and 0.05. We see that using the less strict threshold $\epsilon = 0.05$, the true model is recovered the majority of times across all simulation settings. We also observe that as the number of nodes N increases, the proportion of times that the true model is recovered increases considerably. On the other hand, there is no clear relationship between the number of hyperedges M and the proportion of successful recovery of the true model.

5 Applications

5.1 Star Wars Movie Scenes

Our first application is modeling co-appearance of the main characters in the scenes of the movie “Star Wars: A New Hope”. We collected the scripts of the movie from the Internet Movie Script Database¹ and constructed a hypergraph for the eight main characters so that each character is a vertex in the hypergraph. We define each scene in the movie as a hyperedge with a total of 178 hyperedges, and a character is contained in the scene if he/she speaks in the scene.

We determine the optimal number of clusters and additional clusters using BIC where the results are provided in Table 6. The ELCA model with 3 clusters and 2 additional clusters has the lowest BIC value and is selected. It is worth noting that the standard LCA with 3 clusters is also competitive based on the BIC.

The results from fitting the ELCA model with $G = 3$ and $K = 2$ are provided in Tables 7 and 8. We can see the variation in the size of hyperedges from the parameter estimates \hat{a} and $\hat{\tau}$ with the majority (81%) of hyperedges having size much smaller than the rest of the hyperedges. Thus, one can deduce that a small proportion of the movie scenes have far more characters.

The estimates $\hat{\phi}$ in Table 8 reveal interesting clustering structure for the 8 main characters in the movie. For example, the lead character “Luke” has a strong tendency to appear in the two largest clusters. On the other hand, it is extremely unlikely for “Obi-Wan” and “Han” appear in the same scene.

The estimated primary cluster assignment probabilities from the EM algorithm for each movie scene in the Star Wars movie are shown in chronological order in Fig. 3. We can see from the plot that scenes in the early part of the movie are mainly associated with cluster 1, while cluster 2 contains most of the scenes from roughly scene 40 to scene 100. We can deduce from this, for example, that the character “Han” is very active in the middle part of the movie. On the other hand, there does not appear to be any obvious pattern for the third cluster. The clustering for many early and late movie scenes is relatively uncertain, as shown in the plot.

¹ Movie script data freely available at <https://www.imsdb.com/>.

Table 1 Convergence analysis of the EM algorithm for the ELCA model with 2 primary clusters and 2 additional clusters

Model	M	ϕ	α	π	τ	mi_{s1}	mi_{s2}
10-node ($\pi = (1/2, 1/2), \tau = (1/2, 1/2)$)	100	0.0465	0.0224	0.0269	0.0630	0.0412	0.1561
	500	0.0205	0.0075	0.0083	0.0315	0.0374	0.1463
	1000	0.0124	0.0043	0.0064	0.0199	0.0379	0.1428
10-node ($\pi = (2/3, 1/3), \tau = (1/2, 1/2)$)	100	0.0549	0.0292	0.0147	0.0491	0.0293	0.1450
	500	0.0248	0.0108	0.0082	0.0296	0.0266	0.1454
	1000	0.0209	0.0046	0.0039	0.0199	0.0273	0.1453
10-node ($\pi = (1/2, 1/2), \tau = (2/3, 1/3)$)	100	0.0546	0.0176	0.0173	0.0435	0.0380	0.1332
	500	0.0257	0.0053	0.0106	0.0220	0.0374	0.1328
	1000	0.0146	0.0027	0.0053	0.0173	0.0362	0.1312
10-node ($\pi = (2/3, 1/3), \tau = (2/3, 1/3)$)	100	0.0698	0.0137	0.0213	0.0441	0.0365	0.1430
	500	0.0247	0.0082	0.0094	0.0189	0.0372	0.1279
	1000	0.0168	0.0040	0.0082	0.0132	0.0358	0.1235
20-node ($\pi = (1/2, 1/2), \tau = (1/2, 1/2)$)	100	0.0559	0.0120	0.0216	0.0195	0.0065	0.0750
	500	0.0170	0.0039	0.0102	0.0103	0.0059	0.0720
	1000	0.0114	0.0037	0.0051	0.0101	0.0056	0.0701
20-node ($\pi = (2/3, 1/3), \tau = (1/2, 1/2)$)	100	0.0450	0.0127	0.0250	0.0301	0.0102	0.0640
	500	0.0232	0.0041	0.0080	0.0087	0.0061	0.0620
	1000	0.0112	0.0024	0.0054	0.0082	0.0062	0.0624

Table 1 continued

Model	M	ϕ	a	π	τ	mis_1	mis_2
20-node ($\pi = (1/2, 1/2), \tau = (2/3, 1/3)$)	100	0.0389	0.0120	0.0278	0.0309	0.0090	0.0635
	500	0.0242	0.0040	0.0081	0.0133	0.0089	0.0613
	1000	0.0135	0.0018	0.0077	0.0112	0.0086	0.0604
20-node ($\pi = (2/3, 1/3), \tau = (2/3, 1/3)$)	100	0.0558	0.0100	0.0172	0.0304	0.0082	0.0724
	500	0.0194	0.0039	0.0139	0.0121	0.0068	0.0686
	1000	0.0108	0.0021	0.0071	0.0061	0.0067	0.0627

The ℓ_2 distance between the true parameters of ϕ, a, π, τ and the estimated ones, and the misclassification rates for both the primary (mis_1) and additional clusters (mis_2) are presented

Table 2 Convergence analysis of the EM algorithm for the EICA model with 3 primary clusters and 2 additional clusters

Model	M	ϕ	α	π	τ	mi_{s1}	mi_{s2}
10-node ($\pi = (1/3, 1/3, 1/3)$, $\tau = (1/2, 1/2)$)	100	0.1286	0.0399	0.0235	0.0778	0.1997	0.1858
	500	0.0747	0.0076	0.0108	0.0352	0.1758	0.1692
	1000	0.0541	0.0069	0.0099	0.0138	0.1575	0.1553
10-node ($\pi = (1/2, 1/4, 1/4)$, $\tau = (1/2, 1/2)$)	100	0.1317	0.0368	0.0589	0.0590	0.1715	0.1620
	500	0.0850	0.0117	0.0448	0.0363	0.1582	0.1573
	1000	0.0534	0.0052	0.0216	0.0173	0.1529	0.1542
10-node ($\pi = (1/3, 1/3, 1/3)$, $\tau = (3/4, 1/4)$)	100	0.1329	0.0432	0.0277	0.0522	0.2335	0.1375
	500	0.1053	0.0106	0.0126	0.0160	0.2172	0.1318
	1000	0.0698	0.0063	0.0112	0.0171	0.2038	0.1291
10-node ($\pi = (1/2, 1/4, 1/4)$, $\tau = (3/4, 1/4)$)	100	0.1318	0.0390	0.0782	0.0319	0.2162	0.1485
	500	0.0866	0.0091	0.0521	0.0162	0.1941	0.1292
	1000	0.0745	0.0052	0.0368	0.0158	0.1877	0.1241
20-node ($\pi = (1/3, 1/3, 1/3)$, $\tau = (1/2, 1/2)$)	100	0.1083	0.0194	0.0208	0.0390	0.1655	0.1105
	500	0.0523	0.0039	0.0058	0.0139	0.1293	0.1045
	1000	0.0356	0.0019	0.0028	0.0069	0.1208	0.1014
20-node ($\pi = (1/2, 1/4, 1/4)$, $\tau = (1/2, 1/2)$)	100	0.1217	0.0169	0.0597	0.0398	0.1647	0.1020
	500	0.0618	0.0062	0.0271	0.0182	0.1176	0.0992
	1000	0.0339	0.0027	0.0139	0.0078	0.1094	0.0967

Table 2 continued

Model	M	ϕ	a	π	τ	mis_1	mis_2
20-node ($\pi = (1/3, 1/3, 1/3)$, $\tau = (3/4, 1/4)$)	100	0.1079	0.0205	0.0290	0.0389	0.2275	0.0915
	500	0.0672	0.0083	0.0104	0.0229	0.1728	0.0862
	1000	0.0434	0.0041	0.0038	0.0131	0.1574	0.0807
20-node ($\pi = (1/2, 1/4, 1/4)$, $\tau = (3/4, 1/4)$)	100	0.1265	0.0604	0.0703	0.0389	0.1982	0.0880
	500	0.0724	0.0192	0.0384	0.0207	0.1617	0.0752
	1000	0.0366	0.0025	0.0121	0.0119	0.1426	0.0713

The ℓ_2 distance between the true parameters of ϕ , a , π , τ and the estimated ones, and the misclassification rates for both the primary (mis_1) and additional clusters (mis_2) are presented

Table 3 Percentage of times the lowest BIC values occurred in each model

G	K	N = 10			N = 20			N = 40			N = 500		
		M = 50			M = 50			M = 50			M = 500		
		BIC	BIC	BIC	BIC	BIC							
1	1	0	0	0	0	0	0	0	0	0	0	0	
1	2	0	0	0	0	0	0	0	0	0	0	0	
2	1	52	28	11	26	14	9	12	0	0	0	0	
2	2	42	67	74	67	82	91	88	100	83	0	0	
3	1	4	2	8	0	0	0	0	0	0	0	0	
3	2	2	3	7	7	4	0	0	0	0	0	17	
4	1	0	0	0	0	0	0	0	0	0	0	0	
4	2	0	0	0	0	0	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	0	0	0	0	0	
5	2	0	0	0	0	0	0	0	0	0	0	0	

For the first two columns (Column ‘G’ and ‘K’): bold indicates the true model. For the rest of the columns, the largest values are bolded

Table 4 Percentage of times the lowest BIC values occurred in each model

G	K	N = 10			N = 20			N = 40			N = 500			
		M = 50		M = 200	M = 50		M = 200	M = 50		M = 200	M = 500		M = 200	M = 500
		BIC		BIC	BIC		BIC	BIC	BIC		BIC	BIC	BIC	BIC
1	1	0	0	0	0	0	0	0	0	0	0	0	0	
1	2	0	0	0	0	0	0	0	0	0	0	0	0	
2	1	43	12	2	27	4	0	0	13	0	0	0	0	
2	2	28	13	1	42	9	0	0	58	2	0	0	0	
3	1	14	29	19	9	14	7	0	0	0	0	0	0	
3	2	15	46	78	22	73	84	29	98	100	0	0	0	
4	1	0	0	0	0	0	6	0	0	0	0	0	0	
4	2	0	0	0	0	0	3	0	0	0	0	0	0	
5	1	0	0	0	0	0	0	0	0	0	0	0	0	
5	2	0	0	0	0	0	0	0	0	0	0	0	0	

For the first two columns (Column ‘G’ and ‘K’): bold indicates the true model. For the rest of the columns, the largest values are bolded

Table 5 Proportion of times that the true model can be recovered

True model	Fitted model	N	M	RR ($\epsilon = 0.01$)	RR ($\epsilon = 0.05$)
$G = 2, K = 1$	$G = 2, K = 2$	10	50	0.55	0.83
			100	0.57	0.83
			500	0.66	0.91
$G = 2, K = 1$	$G = 2, K = 2$	20	50	0.64	0.85
			100	0.66	0.83
			500	0.72	0.88
$G = 2, K = 1$	$G = 2, K = 3$	10	50	0.32	0.51
			100	0.27	0.59
			500	0.33	0.62
$G = 2, K = 1$	$G = 2, K = 3$	20	50	0.56	0.78
			100	0.55	0.86
			500	0.59	0.83
$G = 3, K = 1$	$G = 3, K = 2$	10	50	0.55	0.77
			100	0.53	0.80
			500	0.52	0.78
$G = 3, K = 1$	$G = 3, K = 2$	20	50	0.70	0.93
			100	0.67	0.92
			500	0.66	0.89
$G = 3, K = 1$	$G = 3, K = 3$	10	50	0.43	0.65
			100	0.36	0.63
			500	0.33	0.64
$G = 3, K = 1$	$G = 3, K = 3$	20	50	0.54	0.71
			100	0.49	0.76
			500	0.53	0.75

The recovery rates corresponding to $\epsilon = 0.01$ and $\epsilon = 0.05$ for each simulation setting are shown

Table 6 Model selection for the Star Wars data set

No. of clusters	No. of Additional clusters	BIC
1	1	1298.08
1	2	1437.86
2	1	1269.11
2	2	1271.55
3	1	1270.46
3	2	1266.42
3	3	1280.81
4	1	1273.54
4	2	1284.68
5	1	1307.05
5	2	1298.11
5	3	1306.50

The smallest value is bolded

Table 7 Estimates of π , τ and a from fitting the ELCA model with 3 clusters and 2 additional clusters for the Star Wars data set

$\hat{\pi}$	(0.40, 0.40, 0.20)
$\hat{\tau}$	(0.81, 0.19)
\hat{a}	(0.41, 1.00)

Table 8 Estimates of $\{\phi_{ig}\}$ from fitting the ELCA model with 3 clusters and 2 additional clusters for the Star Wars data set

Character	Cluster 1	Cluster 2	Cluster 3
Wedge	0.18	0.00	0.36
Han	0.00	1.00	0.00
Luke	1.00	1.00	0.00
C-3PO	0.75	0.30	0.00
Obi-Wan	0.00	0.00	1.00
Leia	0.12	0.48	0.07
Biggs	0.31	0.00	0.28
Darth Vader	0.19	0.35	0.06

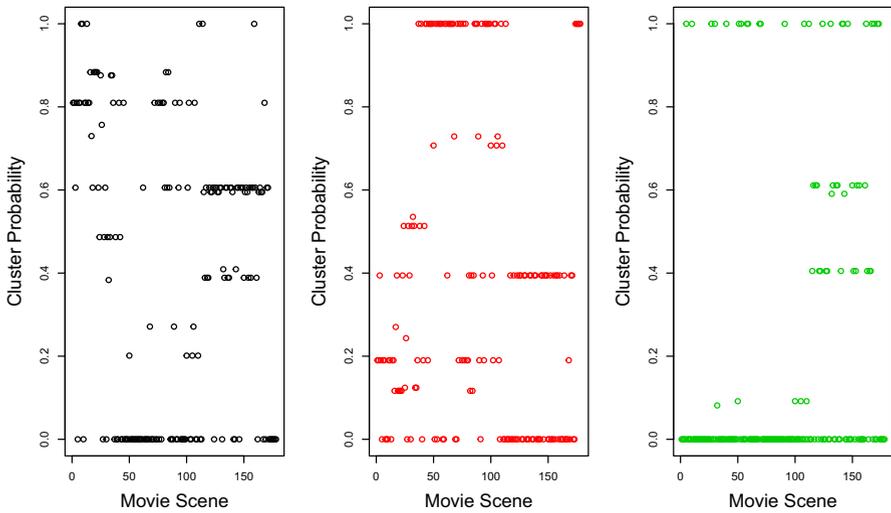


Fig. 3 Probability of primary clusters for movie scenes in Star Wars data set plotted against movie scene number for the ELCA model with 3 primary clusters and 2 additional clusters. Cluster 1 is associated with scenes in the first half of the movie, whereas cluster 2 contains scenes mostly in the middle of the movie. On the other hand, scenes occurring in the second half of the movie are slightly more likely to be associated with cluster 3 compared to scenes occurring in the first half of the movie

The uncertainties in primary clustering are also illustrated in a ternary plot in Fig. 4. Each dot in the plot represents a movie scene, and the three corners of the plot represent the three clusters. The closer the dot is to the corner, the higher probability that the corresponding movie scene belongs to the corresponding cluster. The ternary plot in Fig. 4 shows significant uncertainties in clustering a number of movie scenes into the first two clusters. This is reasonable since for a number of actors including the

Fig. 4 Ternary plot of the *a posteriori* group membership probabilities for the scenes in the Star Wars data set

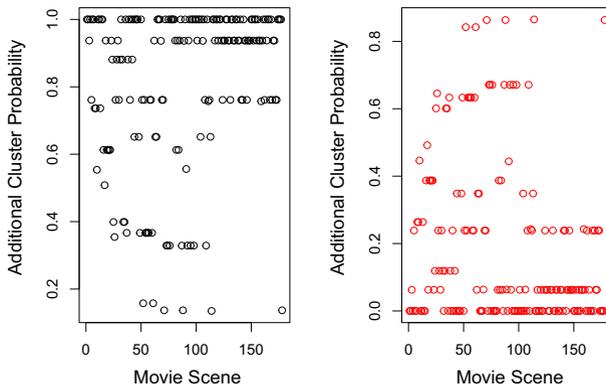
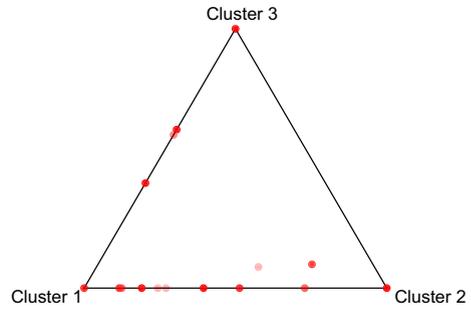


Fig. 5 Probability of additional clusters for movie scenes in Star Wars data set plotted against movie scene number for the ELCA model with 3 primary clusters and 2 additional clusters. Majority of movie scenes are in cluster 1 whereas very few scenes are in cluster 2

lead actor “Luke”, the probabilities of scene appearance are similar for the first two clusters.

The estimated additional cluster assignment probabilities for each movie scene in the Star Wars movie are shown in chronological order in Fig. 5. We observe that majority of the scenes are assigned additional cluster 1 with only a small number of scenes between scene 40 and 100 assigned to additional cluster 2 where these scenes tend to have more characters.

As a comparison, the results from fitting the standard LCA model with 3 clusters are shown in Tables 9 and 10, and a contingency table comparing the primary clustering structure of the ELCA model and the LCA model are given in Table 11. The contingency table shows a very different clustering structure obtained from fitting the standard LCA model versus the ELCA model. We show the estimated cluster assignment probabilities for each movie scene for the LCA model with 3 clusters in chronological order in Fig. 6. In comparing Fig. 3 with Fig. 6, we see that while primary cluster 2 and 3 for the fitted ELCA model are similar with cluster 2 and 3 for the fitted LCA model, there is significant difference between primary cluster 1 in the ELCA model and cluster 1 in the LCA model.

Table 9 Estimates of π from fitting the LCA model with 3 clusters for the Star Wars data set

$\hat{\pi}$	(0.17, 0.61, 0.22)
-------------	--------------------

Table 10 Estimates of $\{p_{ig}\}$ from fitting the LCA model with 3 clusters for the Star Wars data set

Character	Cluster 1	Cluster 2	Cluster 3
Wedge	0.47	0.00	0.00
Han	0.00	0.40	0.00
Luke	0.23	0.74	0.00
C-3PO	0.00	0.24	0.38
Obi-Wan	0.00	0.00	0.60
Leia	0.00	0.21	0.04
Biggs	0.52	0.02	0.00
Darth Vader	0.00	0.18	0.03

Table 11 Contingency table: ELCA with 3 clusters and 2 additional clusters versus LCA with 3 clusters

	LCA		
ELCA	1	2	3
1	16	47	22
2	0	57	0
3	12	0	24

The difference in the clustering structure between the ELCA model and the LCA model is expected as the ELCA model explicitly captures the variation in the size of hyperedges. In comparison, the LCA model cannot decouple the variation in the size of hyperedges from the primary clustering structure. This is a key advantage of the ELCA model where the underlying structure of the size of the hyperedges can be uncovered. Furthermore, as a constrained version of the LCA model with 6 clusters, the ELCA model with 3 primary clusters and 2 additional clusters is far more parsimonious.

5.2 Reuters News articles

As a second application of the ELCA model, we collected news articles published by Reuters² in January 2020. We analyze the co-appearance relationships among the Group of Eight+Five (G8+5) countries. A hypergraph is constructed by defining each news article as a hyperedge and each country as a vertex. A vertex is contained in a hyperedge if the corresponding country is mentioned in the corresponding news article. News articles that do not mention any of the 13 countries were removed, and the resulting hypergraph contains 1828 hyperedges.

The model with 5 clusters and 2 additional clusters was chosen by the BIC and fitted to the data set. The BIC scores for a range of models are shown in Table 12. It is

² <http://www.reuters.com>.

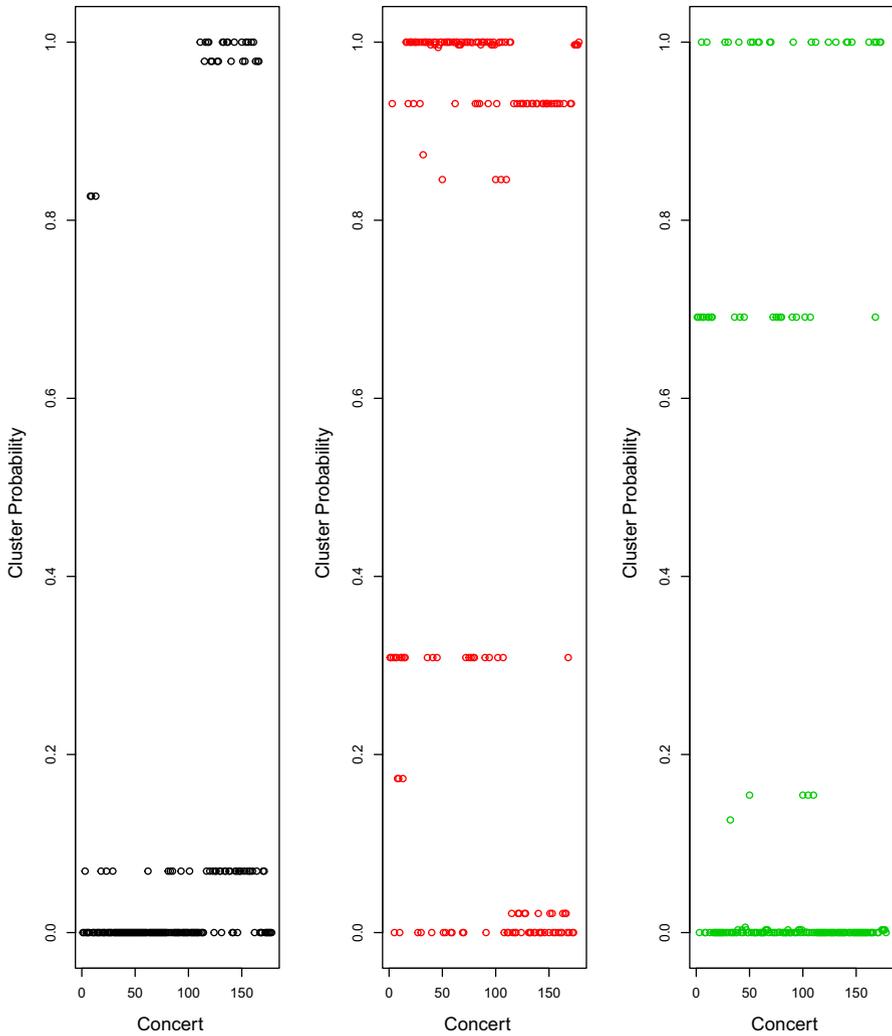


Fig. 6 Probability of clusters for movie scenes in Star Wars data set plotted against movie scene number for the LCA model with 3 clusters. Movie scenes in cluster 1 mostly occurred in the second half of the movie, whereas cluster 2 contains majority of the scenes in the movie. On the other hand, scenes in the first half of the movie are slightly more likely to be associated with cluster 3 compared to scenes in the second half of the movie

worth noting that according to the BIC scores the ELCA models with two additional clusters generally outperform the standard LCA models whereas the standard LCA performs better than the ELCA with three additional clusters.

The parameter estimates $\hat{\pi}$, $\hat{\tau}$ and \hat{a} are given in Table 13. The estimate $\hat{\pi}$ shows that the hyperedges are relatively evenly distributed across the five clusters. We can deduce from \hat{a} and $\hat{\tau}$ that there are a small number of articles mentioning many countries whereas the vast majority of the articles mention very few countries. Specifically,

Table 12 Model selection for Reuters News data set

No. of clusters	No. of additional clusters	BIC
1	1	18,018
1	2	19,005
2	1	17,801
2	2	17,711
2	3	17,723
3	1	17643
3	2	17636
3	3	17652
4	1	17562
4	2	17533
4	3	17625
5	1	17507
5	2	17410
5	3	17611
6	1	17468
6	2	17489
7	1	17514
7	2	17526

The smallest value is bolded

Table 13 Estimates of π , τ and α from fitting the ELCA model with 5 clusters and 2 additional clusters for Reuters News data set

$\hat{\pi}$	(0.16, 0.27, 0.19, 0.12, 0.26)
$\hat{\tau}$	(0.94, 0.06)
$\hat{\alpha}$	(0.28, 1.00)

about 6% of articles mentioned a much larger number of countries compared to the rest of the articles. The incorporation of an additional clustering structure results in significant reduction in the number of parameters.

The clustering structure can be deduced from the estimate $\hat{\phi}$ given in Table 14. China, Russia and USA are among the most popular in articles in cluster 1 whereas China, France and Japan are the most commonly mentioned by articles in cluster 2. Canada, Britain and USA have the highest probability of appearing in articles in cluster 3 whereas Canada, Mexico and USA are the most likely to appear in news articles in cluster 4. Germany, France and Britain are most likely to be mentioned by news articles in cluster 5 (Table 13).

Table 14 Estimates of $\{\phi_{ig}\}$ from fitting the ELCA model with 5 clusters and 2 additional clusters for the Reuters News data set

Country	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
BRA	0.19	0.27	0.00	0.42	0.00
CAN	0.00	0.27	1.00	0.79	0.00
CHN	1.00	1.00	0.46	0.62	0.79
DEU	0.00	0.49	0.38	0.19	0.94
FRA	0.00	0.97	0.80	0.00	1.00
GBR	0.39	0.79	1.00	0.32	1.00
IND	0.66	0.21	0.10	0.45	0.04
ITA	0.00	0.29	0.00	0.13	0.44
JPN	0.12	1.00	0.00	0.00	0.05
MEX	0.00	0.01	0.04	0.95	0.00
RUS	0.95	0.18	0.14	0.10	0.60
USA	1.00	0.35	1.00	1.00	0.47
ZAF	0.20	0.03	0.00	0.04	0.01

The largest three values in each column are bolded

6 Conclusion

We have proposed the Extended Latent Class Analysis model as a generative model for random hypergraphs. Building on a proportionality assumption, the ELCA model introduces two clustering structures for hyperedges which captures variation in the size of hyperedges. The model achieves significant reduction in model complexity compared to the standard Latent Class Analysis model. An EM algorithm has been developed for model fitting where the M-step involves a series of conditional maximization and model selection is performed using BIC. The proposed model is fitted to two data sets and this yields interesting and interpretable structure within the vertices and hyperedges.

Several extensions to the ELCA model are possible. Hyperedges typically have temporal information associated with them, which is the case for the two applications in this paper. Developing a hypergraph model to incorporate such temporal information is of interest. Furthermore, while the ELCA is developed in the context of hypergraph applications, the model could be useful in other applications where the proportionality assumption on latent class conditional probabilities is plausible.

Acknowledgements Funding was provided by Science Foundation Ireland (Grant No. SFI/12/RC/2289-2).

Funding Open Access funding provided by the IReL Consortium

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted

by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A: Proof on Proposition 1

Proof We can write $A = \sum_{i=1}^N A_i$ where $A_i = 1$ if node i appears in the hyperedge and $A_i = 0$ otherwise. Similarly, we write $B = \sum_{i=1}^N B_i$. Let Z_A be the latent cluster assignment of X_A where $Z_A = g$ if X_A is generated from cluster g . Let $Z_B^{(1)}$ and $Z_B^{(2)}$ be the latent cluster and additional clusters assignments of X_B , where $Z_B^{(1)} = g$ and $Z_B^{(2)} = k$ if X_B is generated from cluster g and additional clusters k . We have

$$\begin{aligned}
 E(A) &= \sum_{g=1}^G E(A|Z_A = g) Pr(Z_A = g) \\
 &= \sum_{g=1}^G \sum_{i=1}^N E(A_i|Z_A = g) Pr(Z_A = g) \\
 &= \sum_{g=1}^G \sum_{i=1}^N p_{ig} \pi_g \\
 E(B) &= \sum_{g=1}^G \sum_{k=1}^K E(B|Z_B^{(1)} = g, Z_B^{(2)} = k) Pr(Z_B^{(1)} = g, Z_B^{(2)} = k) \\
 &= \sum_{g=1}^G \sum_{k=1}^K \sum_{i=1}^N E(B_i|Z_B^{(1)} = g, Z_B^{(2)} = k) Pr(Z_B^{(1)} = g, Z_B^{(2)} = k) \\
 &= \sum_{g=1}^G \sum_{k=1}^K \sum_{i=1}^N \phi_{ig} a_k \tau_k \pi_g \\
 &= \sum_{g=1}^G \sum_{i=1}^N p_{ig} \pi_g \\
 &= E(A)
 \end{aligned}$$

For the variance of the LCA model, we have that

$$Var(A) = \sum_{i=1}^N Var(A_i) + 2 \sum_{i < j} Cov(A_i, A_j)$$

where

$$Var(A_i) = E(A_i^2) - E(A_i)^2$$

$$\begin{aligned}
&= Pr(A_i = 1) - Pr(A_i = 1)^2 \\
&= \sum_{g=1}^G p_{ig}\pi_g - \left(\sum_{g=1}^G p_{ig}\pi_g \right)^2 \\
Cov(A_i, A_j) &= E(A_i A_j) - E(A_i)E(A_j) \\
&= Pr(A_i = A_j = 1) - Pr(A_i = 1)Pr(A_j = 1) \\
&= \sum_{g=1}^G p_{ig}p_{jg}\pi_g - \left(\sum_{g=1}^G p_{ig}\pi_g \right) \left(\sum_{g=1}^G p_{jg}\pi_g \right)
\end{aligned}$$

Hence, we have that

$$\begin{aligned}
Var(A) &= \sum_{i=1}^N \sum_{g=1}^G p_{ig}\pi_g - \sum_{i=1}^N \left(\sum_{g=1}^G p_{ig}\pi_g \right)^2 \\
&\quad + 2 \sum_{i < j}^N \sum_{g=1}^G p_{ig}p_{jg}\pi_g - 2 \sum_{i < j}^N \left(\sum_{g=1}^G p_{ig}\pi_g \right) \left(\sum_{g=1}^G p_{jg}\pi_g \right)
\end{aligned}$$

Now,

$$\begin{aligned}
Var(B) &= \sum_{i=1}^N Var(B_i) + 2 \sum_{i < j}^N Cov(B_i, B_j) \\
Var(B_i) &= Pr(B_i = 1) - Pr(B_i = 1)^2 \\
&= \sum_{g=1}^G \sum_{k=1}^K \phi_{ig}a_k\tau_k\pi_g - \left(\sum_{g=1}^G \sum_{k=1}^K \phi_{ig}a_k\tau_k\pi_g \right)^2 \\
&= \sum_{g=1}^G p_{ig}\pi_g - \left(\sum_{g=1}^G p_{ig}\pi_g \right)^2 \\
Cov(B_i, B_j) &= Pr(B_i = B_j = 1) - Pr(B_i = 1)Pr(B_j = 1) \\
&= \sum_{g=1}^G \sum_{k=1}^K \phi_{ig}\phi_{jg}a_k^2\pi_g\tau_k - \left(\sum_{g=1}^G p_{ig}\pi_g \right) \left(\sum_{g=1}^G p_{jg}\pi_g \right)
\end{aligned}$$

We have

$$\begin{aligned}
Var(B) &= \sum_{i=1}^N \sum_{g=1}^G p_{ig}\pi_g - \sum_{i=1}^N \left(\sum_{g=1}^G p_{ig}\pi_g \right)^2 \\
&\quad + 2 \sum_{i < j}^N \sum_{g=1}^G \sum_{k=1}^K \phi_{ig}\phi_{jg}a_k^2\pi_g\tau_k - 2 \sum_{i < j}^N \left(\sum_{g=1}^G p_{ig}\pi_g \right) \left(\sum_{g=1}^G p_{jg}\pi_g \right)
\end{aligned}$$

Now,

$$\begin{aligned} \text{Var}(B) - \text{Var}(A) &= 2 \sum_{i < j}^N \sum_{g=1}^G \sum_{k=1}^K \phi_{ig} \phi_{jg} a_k^2 \pi_g \tau_k - 2 \sum_{i < j}^N \sum_{g=1}^G p_{ig} p_{jg} \pi_g \\ &= 2 \sum_{i < j}^N \sum_{g=1}^G \left(\sum_{k=1}^K \phi_{ig} \phi_{jg} a_k^2 \tau_k - p_{ig} p_{jg} \right) \pi_g \\ &= 2 \sum_{i < j}^N \sum_{g=1}^G \phi_{ig} \phi_{jg} \left(\sum_{k=1}^K a_k^2 \tau_k - \left(\sum_{k=1}^K a_k \tau_k \right)^2 \right) \pi_g \end{aligned}$$

To show the quantity above is non-negative, we have to show that

$$\sum_{k=1}^K a_k^2 \tau_k - \left(\sum_{k=1}^K a_k \tau_k \right)^2 \geq 0$$

which follows from Jensen’s inequality. □

Appendix B: Simulation studies

Model parameters for Tables 1 and 3

$$\begin{aligned} a &= (0.5, 1) \\ \phi &= \begin{pmatrix} 0.8 \cdots 0.8 & 0.1 \cdots 0.1 \\ 0.1 \cdots 0.1 & 0.8 \cdots 0.8 \end{pmatrix} \end{aligned}$$

Model parameters for Tables 2 and 4

$$\begin{aligned} a &= (0.5, 1) \\ \phi &= \begin{pmatrix} 0.8 \cdots 0.8 & 0.1 \cdots 0.1 \\ 0.1 \cdots 0.1 & 0.8 \cdots 0.8 \\ 0.4 \cdots 0.4 & 0.4 \cdots 0.4 \end{pmatrix} \end{aligned}$$

References

Agarwal S, Branson K, Belongie S (2006) Higher order learning with graphs. In: Proceedings of the 23rd international conference on machine learning, ICML’06. Association for Computing Machinery, New York, NY, USA, pp 17–24

Aitkin M, Vu D, Francis B (2014) Statistical modelling of the group structure of social networks. *Soc Netw* 38:74–87

Azondekon R, Harper ZJ, Agossa FR, Welzig CM, McRoy S (2018) Scientific authorship and collaboration network analysis on malaria research in Benin: papers indexed in the Web of Science (1996–2016). *Glob Health Res Policy* 3:11

Borgatti SP, Everett MG (1997) Network analysis of 2-mode data. *Soc Netw* 19:243–269

- Bu J, Tan S, Chen C, Wang C, Wu H, Zhang L, He X (2010) Music recommendation by unified hypergraph: combining social media information and music content. In: Proceedings of the 18th ACM international conference on multimedia, MM' 10. Association for Computing Machinery, New York, NY, USA, pp 391–400
- Celeux G, Govaert G (1991) Clustering criteria for discrete data and latent class models. *J Classif* 8:157–176
- Collins LM, Fidler PL, Wugalter SE, Long JD (1993) Goodness-of-fit testing for latent class models. *Multivar Behav Res* 28:375–389
- de Panafieu É (2015) Phase transition of random non-uniform hypergraphs. *J Discrete Algorithms* 31:26–39
- Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B* 39:1–38
- Doreian P, Batagelj V (2004) Generalized blockmodeling of two-mode network data. *Soc Netw* 6:29–53
- Dyer M, Frieze A, Greenhill C (2015) On the chromatic number of a random hypergraph. *J Combin Theory Ser B* 113:68–122
- Faust K, Willert K, Rowlee D, Skvoretz J (2002) Scaling and statistical models for affiliation networks: patterns of participation among Soviet politicians during the Brezhnev era. *Soc Netw* 24:231–259
- Field S, Frank KA, Schiller K, Riegler-Crumb C, Muller C (2006) Identifying positions from affiliation networks: preserving the duality of people and events. *Soc Netw* 28:97–123
- Friel N, Rastelli R, Wyse J, Raftery AE (2016) Interlocking directorates in Irish companies using a latent space model for bipartite networks. *Proc Natl Acad Sci USA* 113:6629–6634
- Fujimoto K, Chou C-P, Valente TW (2011) The network autocorrelation model using two-mode data: affiliation exposure and potential bias in the autocorrelation parameter. *Soc Netw* 33:231–243
- Goldschmidt C (2005) Critical random hypergraphs: the emergence of a giant set of identifiable vertices. *Ann Probab* 33:1573–1600
- Goodman LA (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215–231
- Handcock MS, Raftery AE, Tantrum JM (2007) Model-based clustering for social networks. *J R Stat Soc Ser A* 170:301–354
- Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network analysis. *J Am Stat Assoc* 97:1090–1098
- Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. *J Am Stat Assoc* 76:33–65
- Hunter DR, Lange K (2004) A tutorial on MM algorithms. *Am Stat* 58:30–37
- Karoński M, Łuczak T (2002) The phase transition in a random hypergraph. *J Comput Appl Math* 142:125–135
- Koskinen J, Edling C (2012) Modelling the evolution of a bipartite network—peer referral in interlocking directorates. *Soc Netw* 34:309–322
- Lange K, Hunter DR, Yang I (2000) Optimization transfer using surrogate objective functions. *J Comput Graph Stat* 9:1–59
- Latapy M, Magnien C, Vecchio ND (2008) Basic notions for the analysis of large two-mode networks. *Soc Netw* 30:31–48
- Latouche P, Birmelé E, Ambroise C (2011) Overlapping stochastic block models with application to the French political blogosphere. *Ann Appl Stat* 5:309–336
- Lazarsfeld PF, Henry NW (1968) Latent structure analysis. Houghton Mifflin, Boston
- Leordeanu M, Sminchisescu C (2012) Efficient hypergraph clustering. In: Lawrence ND, Girolami M (eds) Proceedings of the fifteenth international conference on artificial intelligence and statistics. PMLR, vol 22 of proceedings of machine learning research, La Palma, Canary Islands, pp 676–684
- Li X, Li Y, Shen C, Dick A, Van Den Hengel A (2013) Contextual hypergraph modeling for salient object detection. In: The IEEE international conference on computer vision (ICCV)
- Lind PG, González MC, Herrmann HJ (2005) Cycles and clustering in bipartite networks. *Phys Rev E* 72:66
- Lunagómez S, Mukherjee S, Wolpert RL, Airoidi EM (2017) Geometric representations of random hypergraphs. *J Am Stat Assoc* 112:363–383
- Meng X-L, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80:267–278
- Moody J (2004) The structure of a social science collaboration network: disciplinary cohesion from 1963 to 1999. *Am Sociol Rev* 69:213–238
- Newman MEJ (2001a) Scientific collaboration networks. I. Network construction and fundamental results. *Phys Rev E* 64:016131

- Newman MEJ (2001b) Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys Rev E* 64:016132
- Newman ME (2004) Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim E, Frauenfelder H, Toroczkai Z(eds) *Complex networks*. Springer, Berlin, pp 337–370
- Nowicki K, Snijders TAB (2001) Estimation and prediction for stochastic blockstructures. *J Am Stat Assoc* 96:1077–1087
- Nylund KL, Asparouhov T, Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct Equ Model* 14:535–569
- Perugini S, Gonçalves MA, Fox EA (2004) Recommender systems research: a connection-centric survey. *J Intell Inf Syst* 23:107–143
- Poole D (2015) On the strength of connectedness of a random hypergraph. *Electron J Combin* 22, Paper 1.69, 16
- Purkait P, Chin T, Sadri A, Suter D (2017) Clustering with hypergraphs: the case for large hyperedges. *IEEE Trans Pattern Anal Mach Intell* 39:1697–1711
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6:461–464
- Skvoretz J, Faust K (1999) Logit models for affiliation networks. *Sociol Methodol* 29:253–280
- Snijders TA, Lomi A, Torló VJ (2013) A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. *Soc Netw* 35:265–276
- Stasi D, Sadeghi K, Rinaldo A, Petrovic S, Fienberg S (2014) β models for random hypergraphs with a given degree sequence. In: *Proceedings of COMPSTAT 2014—21st international conference on computational statistics*, pp 593–600
- Wang YH (1993) On the number of successes in independent trials. *Stat Sin* 3:295–312
- Wang P, Sharpe K, Robins G, Pattison P (2009) Exponential random graph (p^*) models for affiliation networks. *Soc Netw* 31:12–25
- Wang P, Pattison P, Robins G (2013) Exponential random graph model specifications for bipartite networks—a dependence hierarchy. *Soc Netw* 35:211–222
- Zhou D, Huang J, Schölkopf B (2007) Learning with hypergraphs: clustering, classification, and embedding. In: Schölkopf B, Platt JC, Hoffman T (eds) *Advances in neural information processing systems* 19. MIT Press, pp 1601–1608

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.