Latent Association Graph Inference for Binary Transaction Data

ISBA World Meeting 2022

Luis Carvalho[1] and David Reynolds[2]

June 2022

[1]Boston University, [2]University of New Hampshire

Frequent Itemset Mining

- **Goal**: given collection of item groups, identify sets of items that are frequently observed together.

Transaction 1	Transaction 2	Transaction 3
Bacon	Baking powder	Beer
Bread	**Bread**	Chips
Cheese	**Eggs**	**Bread**
Eggs	Flour	**Eggs**
Juice	**Milk**	Meat
Milk	Oil	**Milk**

Frequent Itemset Mining

- **Goal**: given collection of item groups, identify sets of items that are frequently observed together.

Transaction 1	Transaction 2	Transaction 3
Bacon	Baking powder	Beer
Bread	**Bread**	Chips
Cheese	**Eggs**	**Bread**
Eggs	Flour	**Eggs**
Juice	**Milk**	Meat
Milk	Oil	**Milk**

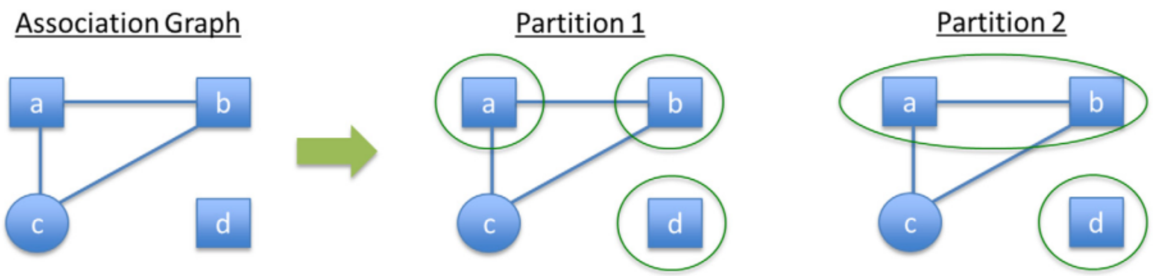
- Applications in databases, bioinformatics, image classification, and market transaction data.

Frequent Itemset Mining

- Most popular approach: *Apriori* algorithm, selecting increasingly large overlapping itemsets
    - Enumerative, so computationally hard
    - Focus on frequency, so also hard to interpret

Frequent Itemset Mining

- Most popular approach: *Apriori* algorithm, selecting increasingly large overlapping itemsets
  - Enumerative, so computationally hard
  - Focus on frequency, so also hard to interpret
- Our approach: use a **latent association graph** (LAG) with a **clique cover** representing transactions



Transaction model

- Given n items and "popularity" parameters γ, a **chordal** LAG G capturing co-purchasing patterns has density

$$P(G) \propto \prod_{u,v \in \{1,\ldots,n\}, u<v} \frac{\exp\{I((u,v) \in G)(\gamma_u + \gamma_v)\}}{1 + \exp\{\gamma_u + \gamma_v\}}$$

and prior $\gamma \sim N(0, cI_n)$, with c large

Transaction model

- Given n items and "popularity" parameters γ, a **chordal** LAG G capturing co-purchasing patterns has density

$$P(G) \propto \prod_{u,v \in \{1,\ldots,n\}, u<v} \frac{\exp\{I((u,v) \in G)(\gamma_u + \gamma_v)\}}{1 + \exp\{\gamma_u + \gamma_v\}}$$

  and prior $\gamma \sim N(0, cI_n)$, with c large

- A transaction T is a disjoint collection of k "cliques":
  - The cardinality follows an *Ewens* distribution with parameter θ:

$$P(k \mid \theta) = \frac{S_n^k \theta^k}{\theta(\theta+1)\cdots(\theta+n-1)},$$

    a model for the number of different *types* of elements in a sample of size n

  - Ewens parameter has quasi-conjugate prior $P(\theta) \propto (\Gamma(\theta)/\Gamma(\theta+n))^{\nu}\theta^{\eta}$ with ν and η small

Transaction model

- A transaction T is a disjoint collection of k "cliques":
    - Each clique $c \in T$ has probability

      $$\pi_c \propto \exp\left(\alpha_{|c|} + |c|^{-1} \sum_{v \in c} \beta_v\right) \doteq \exp(x_c^\top(\alpha, \beta))$$

      where α controls for clique cardinality and β represents item (frequency) popularities, with a joint non-informative prior
    - We set conditions on α to favor small, often *minimum*, clique covers of T, roughly $\alpha_k > \alpha_{k-1} + \rho(\theta)$ where ρ is a penalty for introducing a new clique in the cover

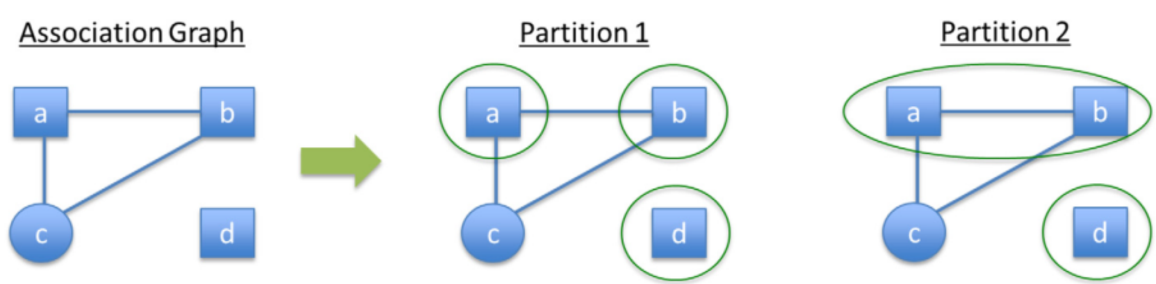
Transaction model

- A transaction T is a disjoint collection of k "cliques":
    - Each clique $c \in T$ has probability

$$\pi_c \propto \exp\left(\alpha_{|c|} + |c|^{-1} \sum_{v \in c} \beta_v\right) \doteq \exp(x_c^\top(\alpha, \beta))$$

      where α controls for clique cardinality and β represents item (frequency) popularities, with a joint non-informative prior

    - We set conditions on α to favor small, often *minimum*, clique covers of T, roughly $\alpha_k > \alpha_{k-1} + \rho(\theta)$ where ρ is a penalty for introducing a new clique in the cover



Finding posterior modes

- Given transaction set $\mathcal{T}$, alternate between finding the conditional posterior modes of $[G, S, k(S) \mid \alpha, \beta, \gamma, \theta, \mathcal{T}]$ and $[\alpha, \beta, \gamma, \theta \mid G, S, k(S), \mathcal{T}]$:
  - **Initialize** G with a minimal triangulation of an inferred graph based on Fisher exact tests for each pair of items; set S and k using minimum clique cover on G
  - **Update** hyper-parameters $\alpha, \beta, \gamma, \theta$ via regularized IRLS as usual for GLMs
  - **Update** G with a specialized neighborhood search in the space of transaction-consistent chordal graphs; again minimum clique cover update for S and k

Finding posterior modes

- Given transaction set $\mathcal{T}$, alternate between finding the conditional posterior modes of $[G, S, k(S) \,|\, \alpha, \beta, \gamma, \theta, \mathcal{T}]$ and $[\alpha, \beta, \gamma, \theta \,|\, G, S, k(S), \mathcal{T}]$:
  - **Initialize** G with a minimal triangulation of an inferred graph based on Fisher exact tests for each pair of items; set S and k using minimum clique cover on G
  - **Update** hyper-parameters $\alpha, \beta, \gamma, \theta$ via regularized IRLS as usual for GLMs
  - **Update** G with a specialized neighborhood search in the space of transaction-consistent chordal graphs; again minimum clique cover update for S and k
- **Cutting corners**: a clique partition s of a transaction has a prohibitively combinatorial normalizing constant, so we adopt an approximation inspired by Breslow's method:

$$
\begin{aligned}
P(s \,|\, k(s), G, \alpha, \beta) &= \frac{\prod_{c \in s} \exp(x_c^\top(\alpha, \beta))}{\sum_{\tilde{s}: k(\tilde{s}) = k(s)} \prod_{\tilde{c} \in \tilde{s}} \exp(x_{\tilde{c}}^\top(\alpha, \beta))} \\
&\approx \frac{\prod_{c \in s} \exp(x_c^\top(\alpha, \beta))}{\left\{ \sum_{\tilde{c} \in C(G)} \exp(x_{\tilde{c}}^\top(\alpha, \beta)) \right\}^{k(s)}}
\end{aligned}
$$

MCMC sampling

- Again we iterate but by Gibbs sampling $[G, S, k(S) \mid \alpha, \beta, \gamma, \theta, \mathcal{T}]$ and $[\alpha, \beta, \gamma, \theta \mid G, S, k(S), \mathcal{T}]$ using Metropolis-Hastings steps
  - **Initialize** chains at the estimated posterior modes (warm start)
  - At iteration t, **sample** candidate G^* from the chordal neighborhood of G^t, then (S^*, k^*) using a randomized perfect elimination scheme (PES)

Acceptance/rejection ratio is approximately

$$\log R([G^*, S^*, k^*], [G^t, S^t, k^t]) \approx \sum_{u,v} \delta_{G^*,G^t}(u,v)(\gamma_u^t + \gamma_v^t) + \sum_{i=1}^{m}(k_i^* - k_i^t)\log\theta^t + \sum_{c \in s_i^*} x_c^\top(\alpha^t, \beta^t) - \sum_{c \in s_i^t} x_c^\top(\alpha^t, \beta^t)$$

with $\delta_{G^*,G^t}(u,v) = I((u,v) \in G^*) - I((u,v) \in G^t)$

MCMC sampling

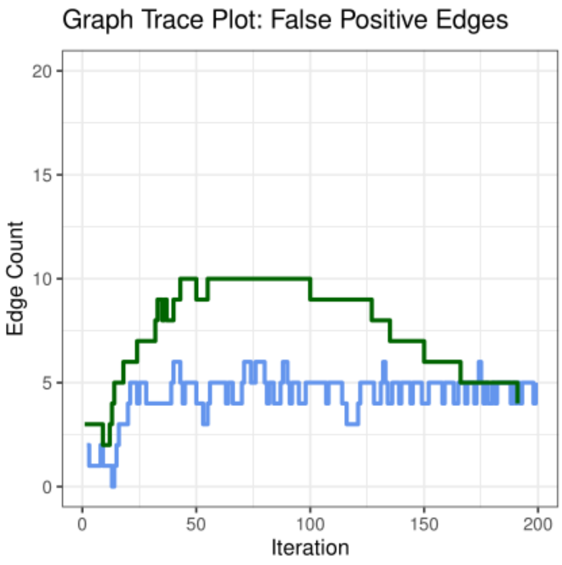
- Again we iterate but by Gibbs sampling $[G, S, k(S) \mid \alpha, \beta, \gamma, \theta, \mathcal{T}]$ and $[\alpha, \beta, \gamma, \theta \mid G, S, k(S), \mathcal{T}]$ using Metropolis-Hastings steps
  - **Initialize** chains at the estimated posterior modes (warm start)
  - At iteration t, **sample** candidate G^* from the chordal neighborhood of G^t, then (S^*, k^*) using a randomized perfect elimination scheme (PES)
  - **Sample** α, β, γ using Metropolis adjusted Langevin algorithm (MALA) steps, as usual in GLMs, but for θ use a tailored gamma proposal

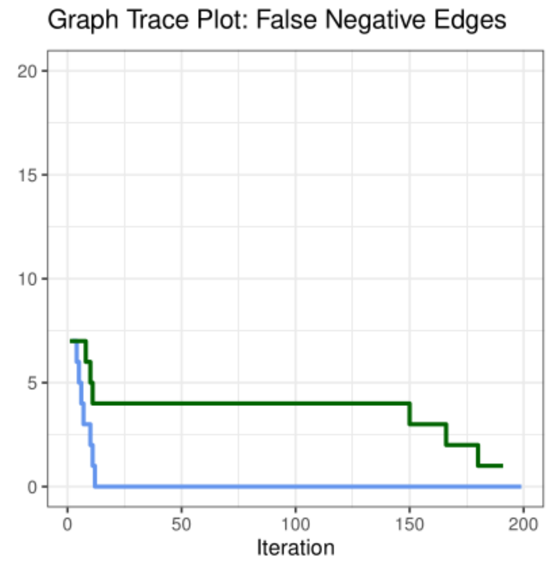
MCMC sampling

- Again we iterate but by Gibbs sampling $[G, S, k(S) \mid \alpha, \beta, \gamma, \theta, \mathcal{T}]$ and $[\alpha, \beta, \gamma, \theta \mid G, S, k(S), \mathcal{T}]$ using Metropolis-Hastings steps
  - **Initialize** chains at the estimated posterior modes (warm start)
  - At iteration t, **sample** candidate G^* from the chordal neighborhood of G^t, then (S^*, k^*) using a randomized perfect elimination scheme (PES)
  - **Sample** α, β, γ using Metropolis adjusted Langevin algorithm (MALA) steps, as usual in GLMs, but for θ use a tailored gamma proposal
- Overall, still computationally expensive but manageable
  - Bottleneck: exploring the space of chordal LAGs and sampling clique covers
  - Optimizing and sampling hyper-parameters is easier, but still based on approximations

Simulation Studies

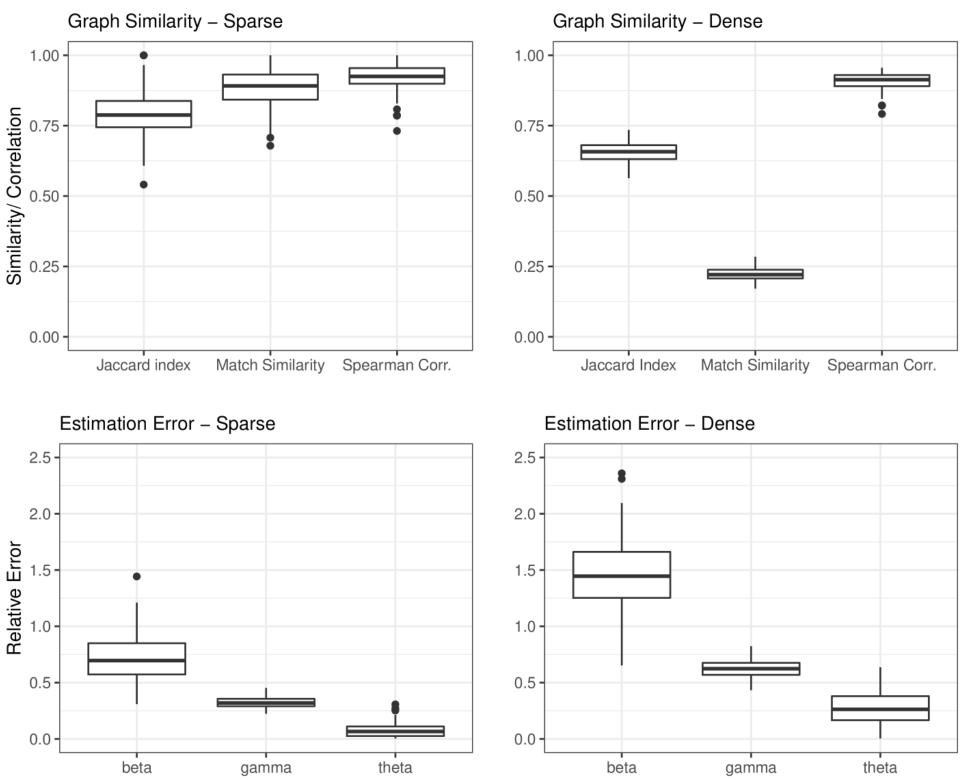
- Generate 100 datasets from $n = 30$ items under two scenarios: sparse, with $\gamma_v \overset{\text{iid}}{\sim} N(-2, 1)$, and dense, with $\gamma_v \overset{\text{iid}}{\sim} N(-1, 1)$, both with $\beta \sim N(0, I_n)$ and $\theta = 0.25$
  - Low false positive and negative rates: e.g. for a LAG with 500 edges, two chains after burn-in:





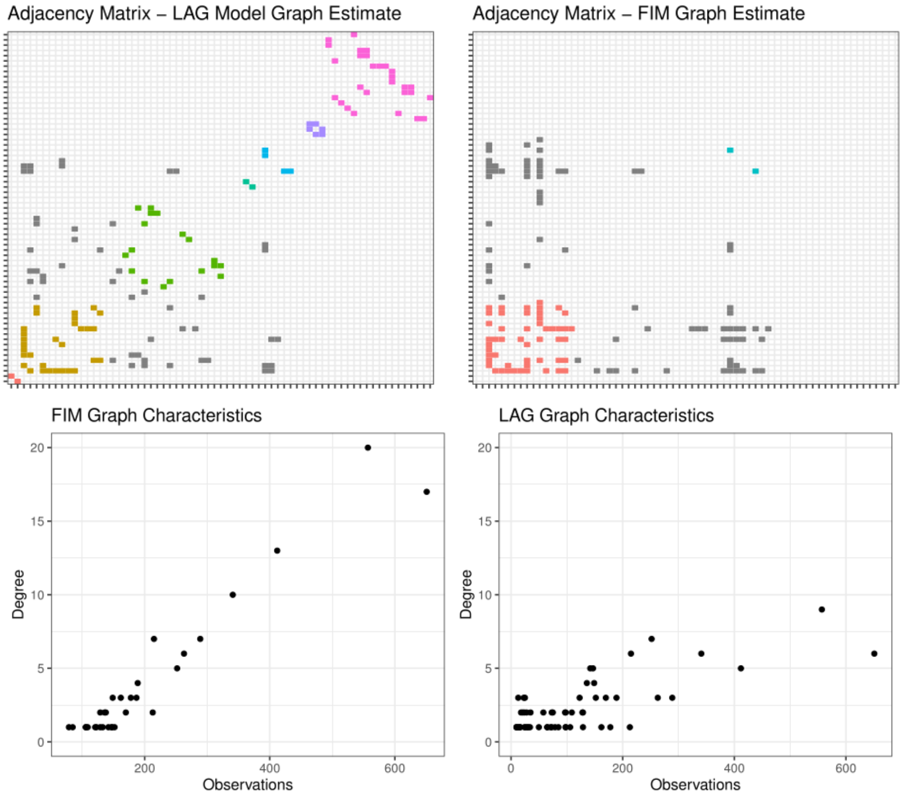
Simulation Studies

- Generate 100 datasets from $n = 30$ items under two scenarios: sparse and dense
  - Low false positive and negative rates
  - Dense LAGs perform worse, due to confounding high β with high γ



Case Study: Instacart

- Random sample of $5{,}000$ transactions with $12{,}114$ items
- Compared to FIM (Apriori): sparser representation, accuracy comparable to *peak* FIM performance but often outperforms FIM on predictive accuracy



Discussion

- LAG is arguably more representative and interpretable model, but there's no free lunch:
  - We had to cut many corners and develop new optimization and sampling routines for chordal graphs
  - Current code is still slow, being implemented in *R*
  - The elephant in the room: MCMC methods for **discrete** parameters

Discussion

- LAG is arguably more representative and interpretable model, but there's no free lunch:
    - We had to cut many corners and develop new optimization and sampling routines for chordal graphs
    - Current code is still slow, being implemented in *R*
    - The elephant in the room: MCMC methods for **discrete** parameters
- Future directions:
    - Develop more efficient sampling procedures
    - Port most of the code to C/C++
    - Apply methodology to other fields, including databases and bioinformatics

Discussion

- LAG is arguably more representative and interpretable model, but there's no free lunch:
  - We had to cut many corners and develop new optimization and sampling routines for chordal graphs
  - Current code is still slow, being implemented in *R*
  - The elephant in the room: MCMC methods for **discrete** parameters
- Future directions:
  - Develop more efficient sampling procedures
  - Port most of the code to C/C++
  - Apply methodology to other fields, including databases and bioinformatics

Reynolds, D. and Carvalho, L., *Computational Statistics and Data Analysis* 160 (2021)

Thanks!