High-dimensional problems in Correlation Mining and Distributed Load Balancing

Miheer Dewaskar

PhD Candidate, Department of Statistics and Operations Research

Advisors: Shankar Bhamidi, Amarjit Budhiraja, and Andrew Nobel.

University of North Carolina at Chapel Hill

16th April, 2021

Two problems that we will consider

Finding Bimodules in high-dimensional multi-view data

Bimodule: group of features from two data matrices that have significant aggregate correlation.

Application to eQTL analysis in genomics: discover SNP-gene association networks.

How to do this in a statistically principled and computationally efficient way?

R software package : https://github.com/miheerdew/cbce.

Joint work with **John Palowitch, Mark He, Michael I. Love, and Andrew B. Nobel.**

Limit theorems for the Supermarket model

Supermarket model: A processing system with multiple queues, where jobs are assigned to queues using a randomized routing scheme.

Motivated by load balancing problem in large data centers, we obtain limit theorems for the Supermarket model as the number of queues increases.

Joint work with **Shankar Bhamidi and Amarjit Budhiraja.**

Two problems that we will consider

Finding Bimodules in high-dimensional multi-view data

Bimodule: group of features from two data matrices that have significant aggregate correlation.

Application to eQTL analysis in genomics: discover SNP-gene association networks.

How to do this in a statistically principled and computationally efficient way?

R software package : https://github.com/miheerdew/cbce.

Joint work with **John Palowitch, Mark He, Michael I. Love, and Andrew B. Nobel.**

Limit theorems for the Supermarket model

Supermarket model: A processing system with multiple queues, where jobs are assigned to queues using a randomized routing scheme.

Motivated by load balancing problem in large data centers, we obtain limit theorems for the Supermarket model as the number of queues increases.

Joint work with **Shankar Bhamidi and Amarjit Budhiraja.**

Context for the Supermarket model: data centers

Figure: Google data center in Eemshaven, Netherlands (`google.com/about/datacenters/`)

The Supermarket model with parameters (λ, d, n)

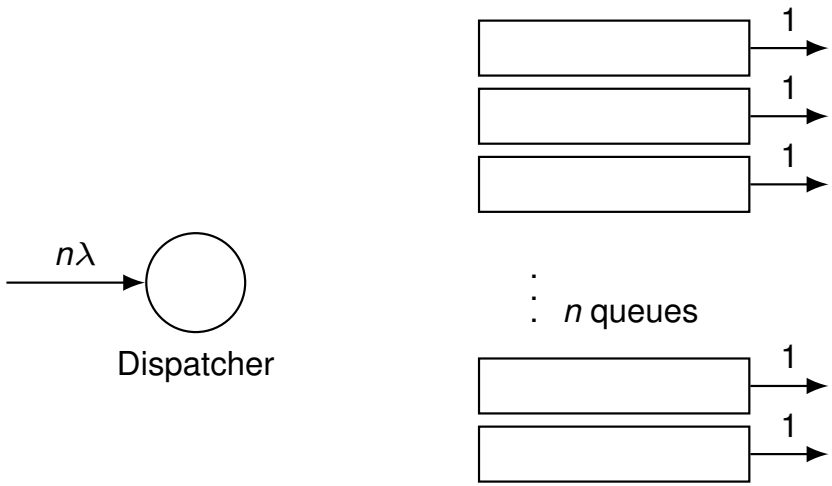


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

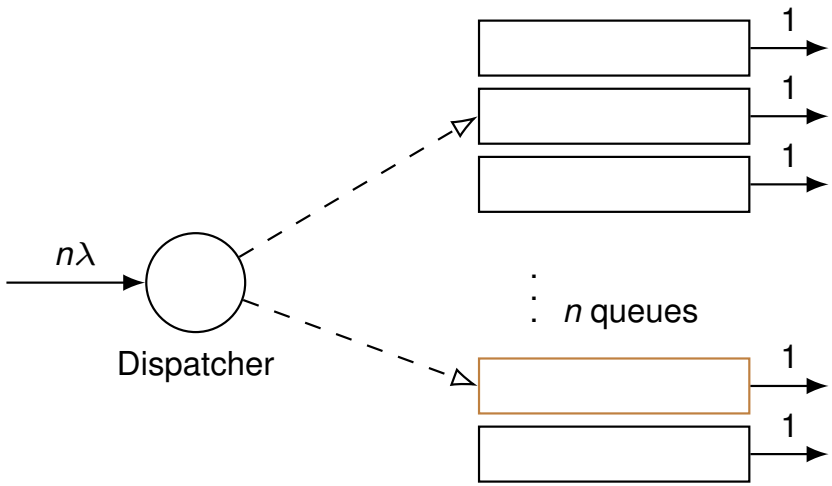


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

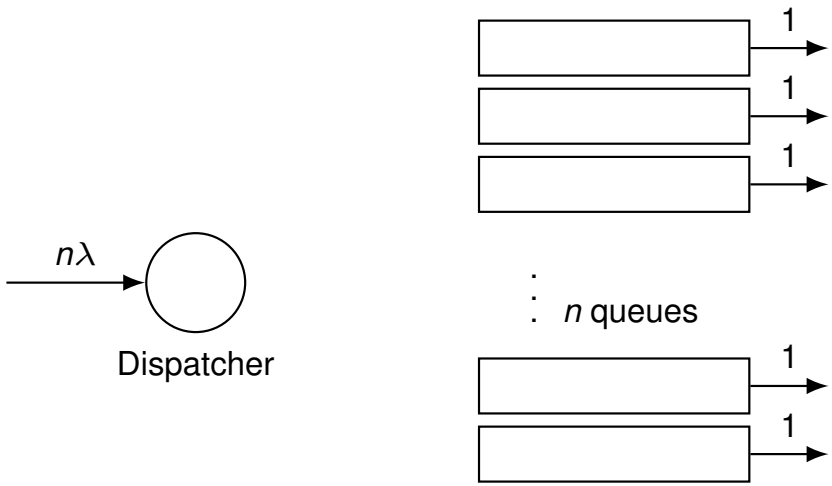


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

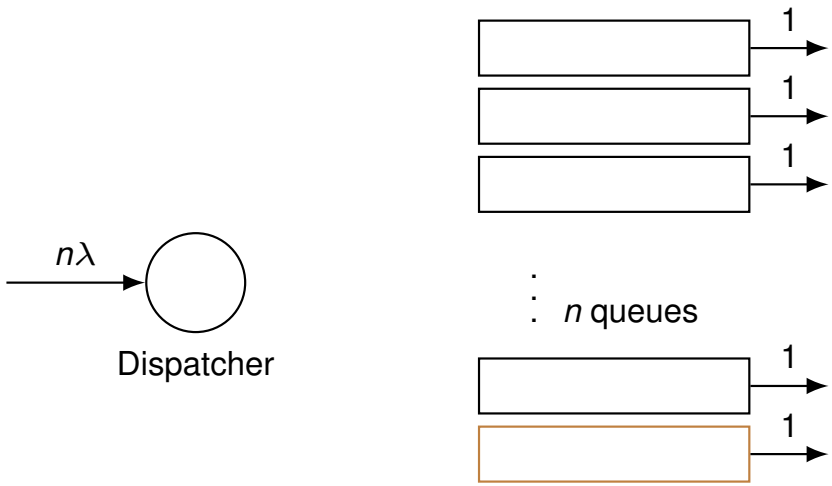


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

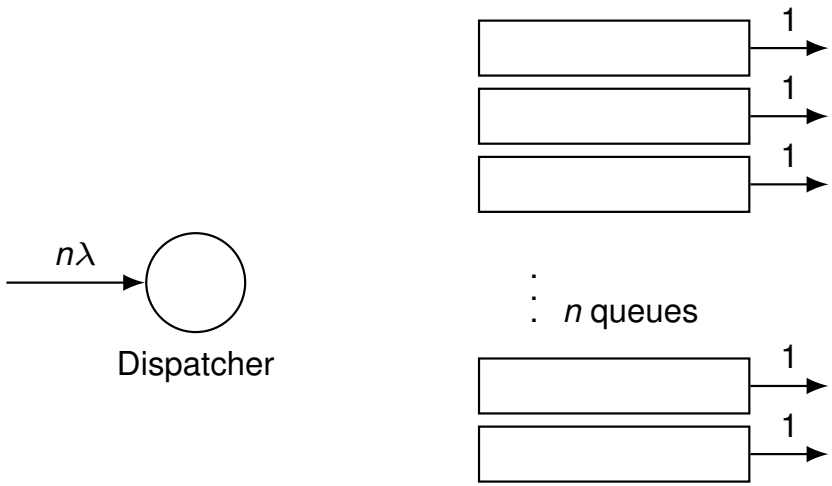


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

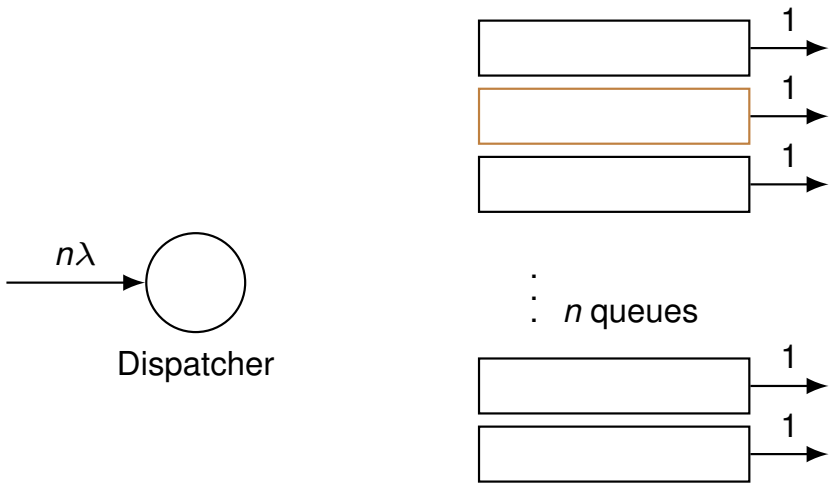


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

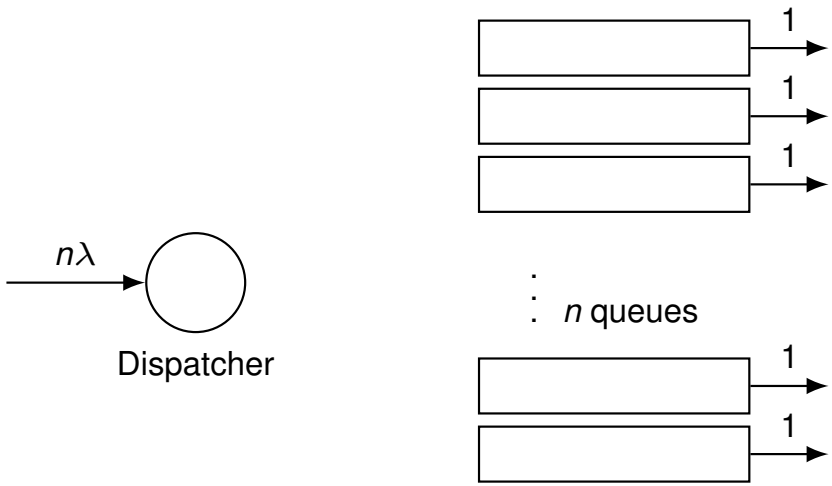


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

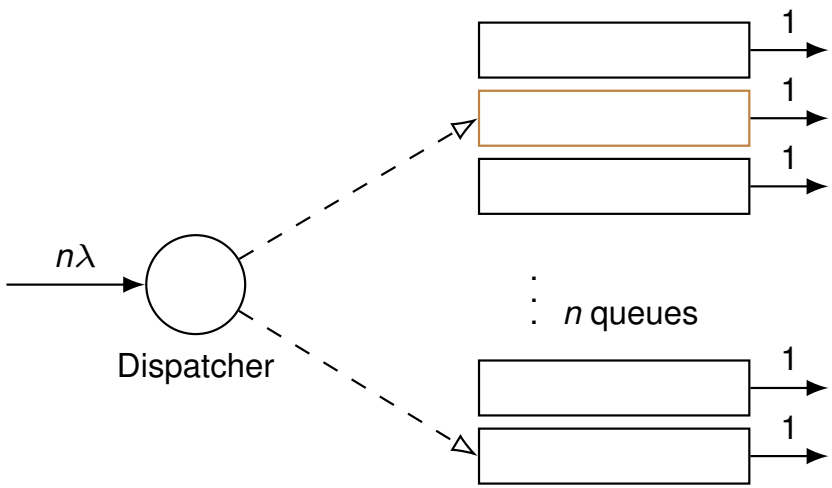


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

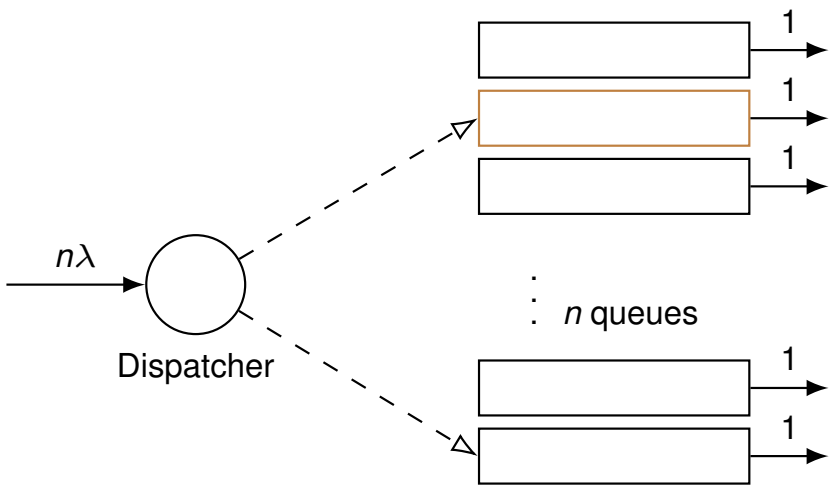


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

The Supermarket model with parameters (λ, d, n)

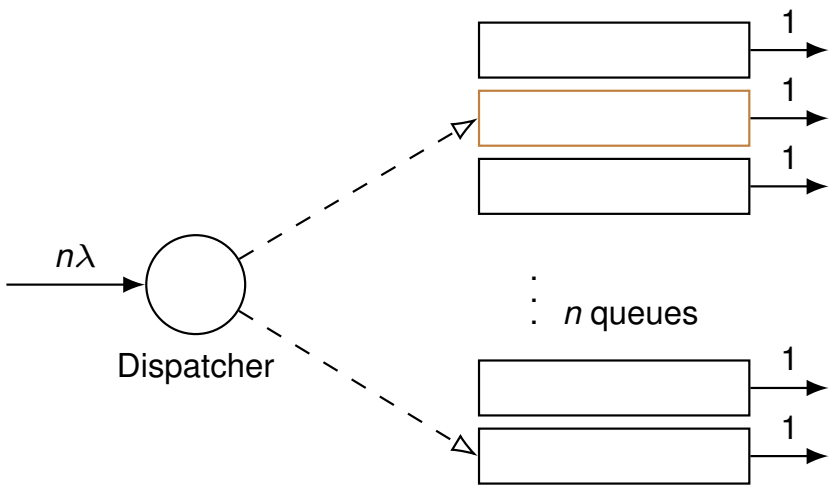


Figure: Supermarket model with n queues and $d = 2$.

- # of choices: $d \in \{1, \ldots, n\}$
- Communication per job: d
- Compare $d = 1$ vs. $d = n$.

"Power of choice"

$d = 2$ is much better than $d = 1$

Analysis for fixed $\lambda < 1$ and $d \geq 2$, as $n \uparrow \infty$.

Vvedenskaya, Dobrushin and Karpelevich 1996; Mitzenmacher 2001; Luczak and McDiarmid 2006.

Interest in heavy traffic: $\lambda = \lambda_n \uparrow 1$

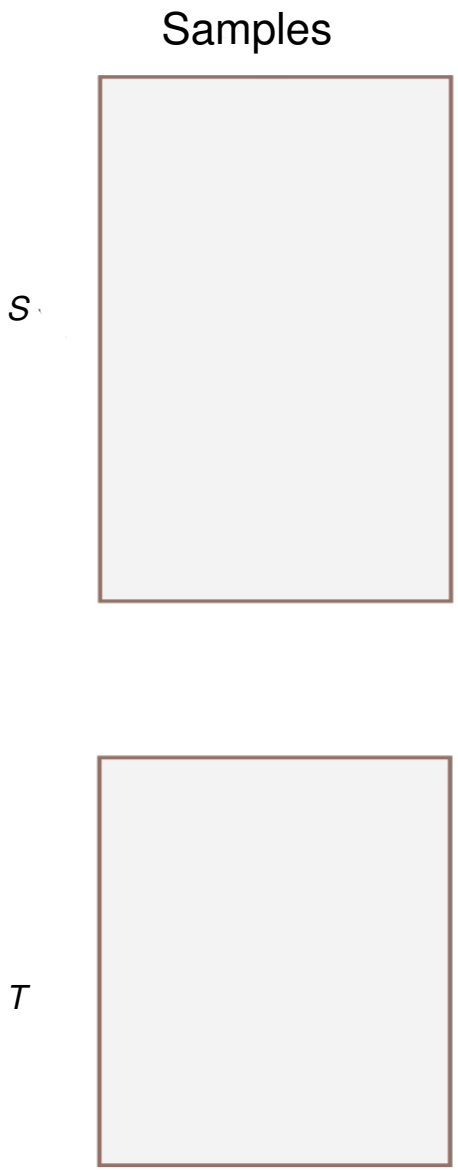
- $d = n$ can achieve optimal load balancing as $\lambda_n \uparrow 1$ and $n \uparrow \infty$.
- $d = O(1)$ cannot achieve optimal load balancing.

Eschenfeldt and Gamarnik (2016, 2018)

As $n \uparrow 1$ and $\lambda_n \uparrow 1$, we prove functional LLN and CLT when $d = d_n$ satisfies $1 \ll d_n \leq n$.

Finding Bimodules in multi-view data

Multi-view data and a related exploratory problem



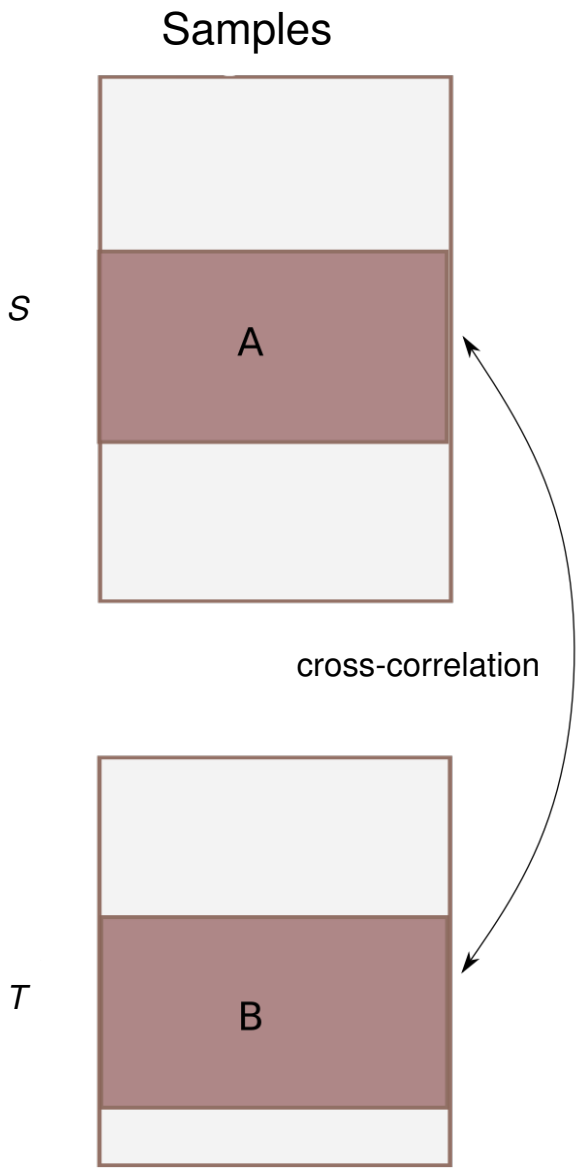
Measurements of two types of features
$S = \{s_1, \ldots, s_p\}$ & $T = \{t_1, \ldots, t_q\}$
on n common samples. Typically $p, q \geq n$.

Examples

- Samples are temporal measurements from
  $S = \{p$ temperature stations$\}$ and
  $T = \{q$ precipitation stations$\}$ worldwide.
- Taken from diverse habitats, samples measure
  $S = \{p$ environmental features$\}$ and
  $T = \{q$ microbial species$\}$ abundance.

How are features from S and T associated?

Exploratory problem of interest



We distinguish between two types of correlations

cross-correlation (CC) b/w features $s \in S$ and $t \in T$

intra-correlation b/w features $s, s' \in S$ or $t, t' \in T$.

Bimodule (rough definition)

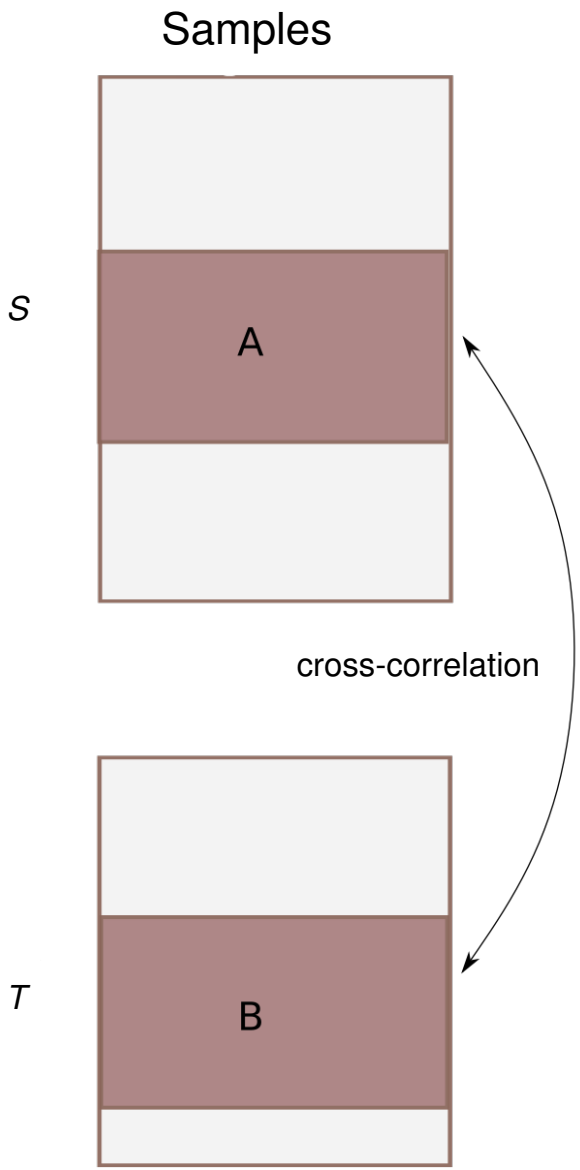
(A, B) is a bimodule if

- $A \subseteq S$ and $B \subseteq T$
- A and B have significant aggregate CC.

Motivation to aggregate CCs

- Capture complex associations between feature groups A and B
- Improve power by amplifying weak signal

Exploratory problem of interest



We distinguish between two types of correlations

cross-correlation (CC) b/w features $s \in S$ and $t \in T$

intra-correlation b/w features $s, s' \in S$ or $t, t' \in T$.

Bimodule (rough definition)

(A, B) is a bimodule if

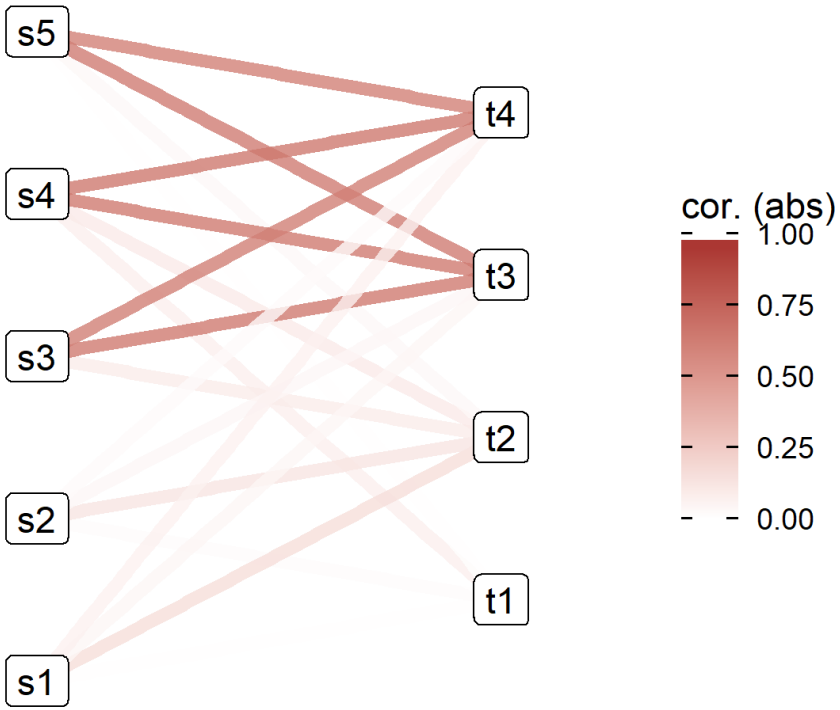
- $A \subseteq S$ and $B \subseteq T$
- A and B have significant aggregate CC.

Motivation to aggregate CCs

- Capture complex associations between feature groups A and B
- Improve power by amplifying weak signal

Bimodules from a network perspective

cross-correlation (CC) networks



Bimodules: communities in this network.

Example: $A = \{s_3, s_4, s_5\}$ and $B = \{t_3, t_4\}$.

Community (rough definition)

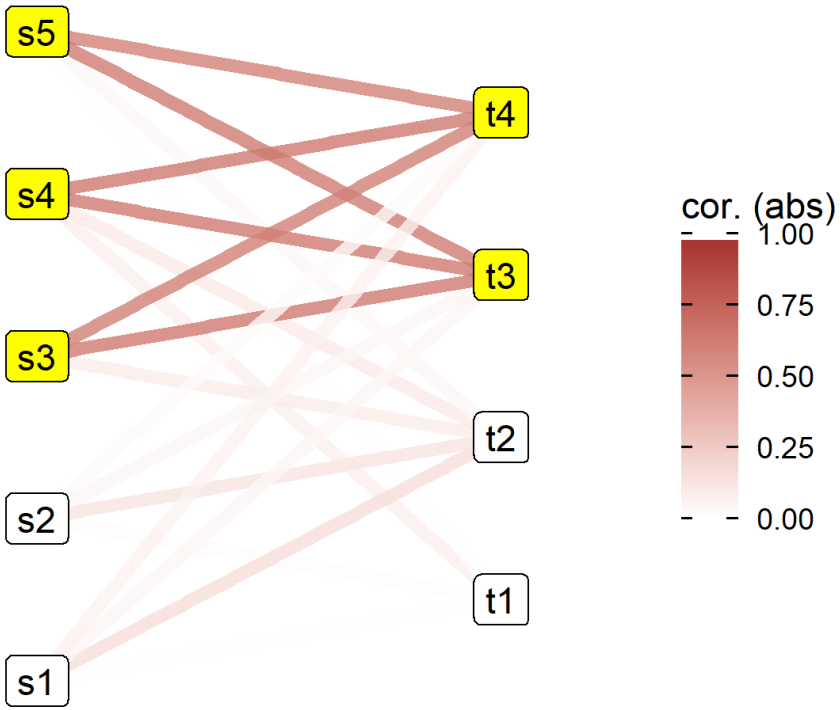
Nodes in a community are more correlated, on average, to nodes inside the community than to nodes outside.

$S = \{s_1, \ldots, s_5\}$, $T = \{t_1, \ldots, t_4\}$

Weights: sample correlation (abs.)

Bimodules from a network perspective

cross-correlation (CC) networks



Bimodules: communities in this network.

Example: $A = \{s_3, s_4, s_5\}$ and $B = \{t_3, t_4\}$.

Community (rough definition)

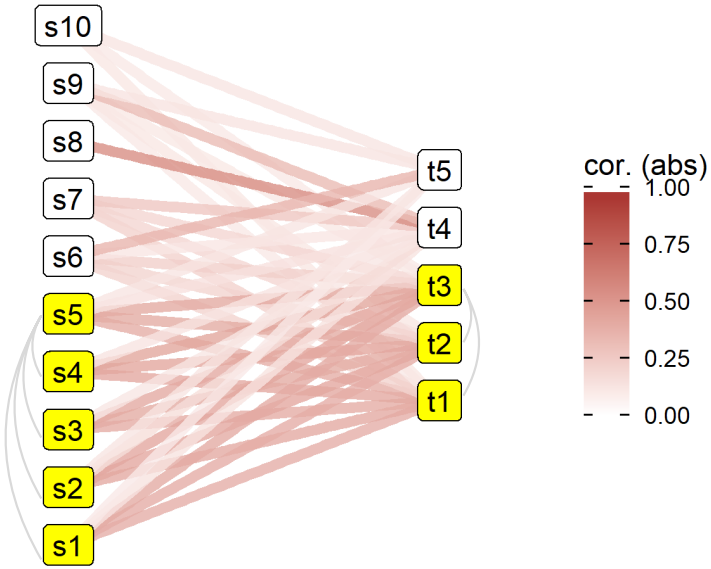
Nodes in a community are more correlated, on average, to nodes inside the community than to nodes outside.

$S = \{s_1, \ldots, s_5\}$, $T = \{t_1, \ldots, t_4\}$

Weights: sample correlation (abs.)

Cross-correlation (CC) matrix may not be sufficient

role of intra-correlations



(A, B) is a community in the CC network.

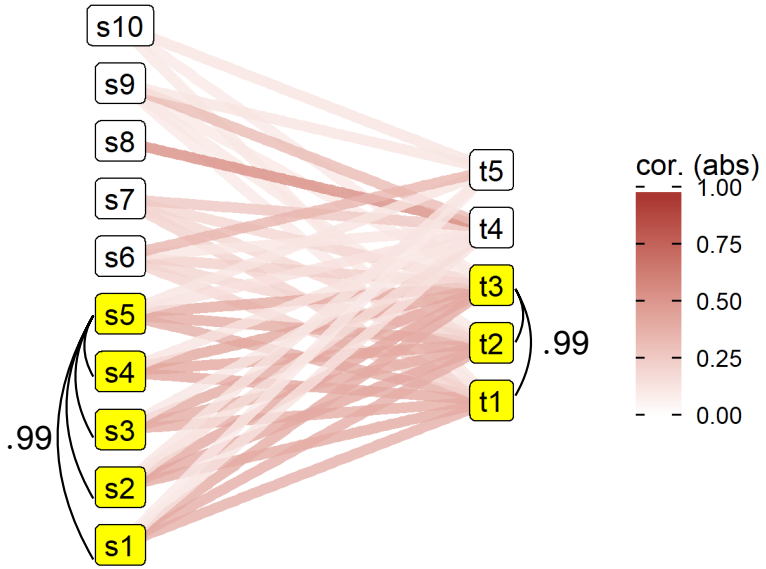
Likely to see this community by chance in random data?

- Depending only on CC can mislead.
- Must account for *intra-correlations* while assessing bimodule significance.

$A = \{s_1, \ldots, s_5\}$, $B = \{t_1, t_2, t_3\}$

Cross-correlation (CC) matrix may not be sufficient

role of intra-correlations



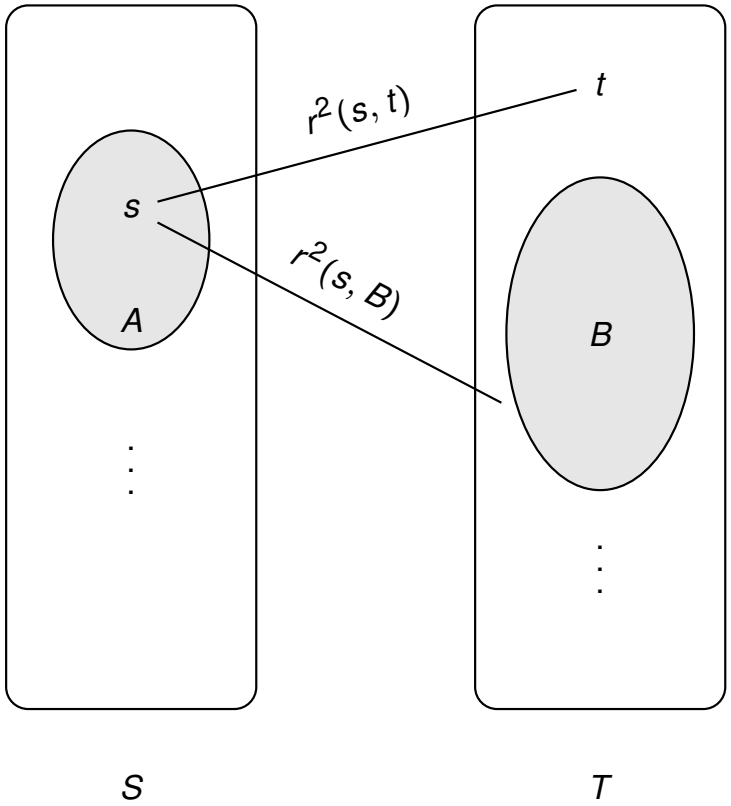
$A = \{s_1, \ldots, s_5\}$, $B = \{t_1, t_2, t_3\}$

(A, B) is a community in the CC network.

Likely to see this community by chance in random data? Yes

- Depending only on CC can mislead.
- Must account for *intra-correlations* while assessing bimodule significance.

Stable Bimodules



Stable bimodule (definition)

(A, B) is a *stable bimodule* if

$$A = \{s \in S \mid r^2(s, B) \text{ is significant}\}, \text{ and}$$

$$B = \{t \in T \mid r^2(A, t) \text{ is significant}\}.$$

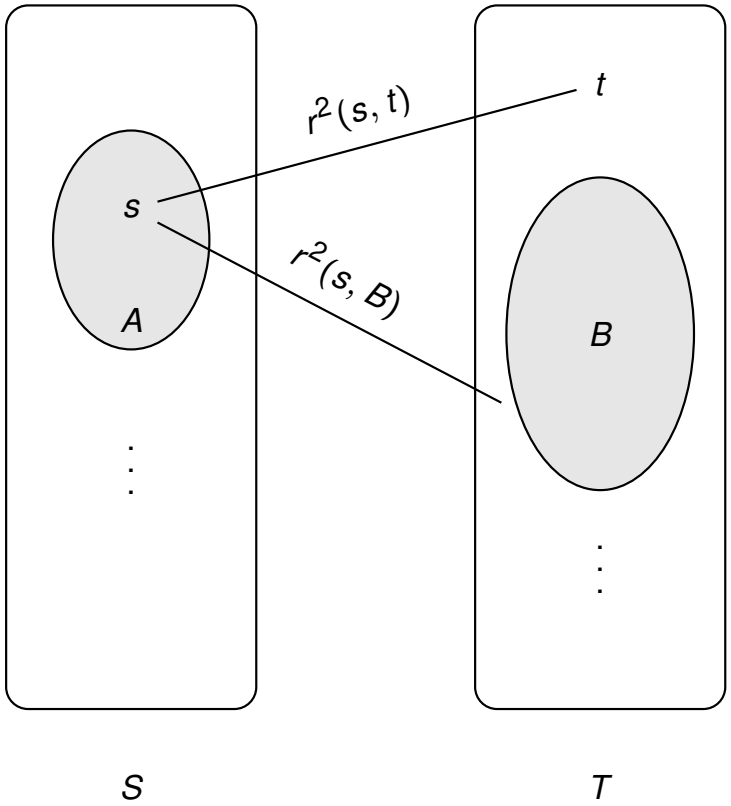
- Recursive definition like a community; made precise using hypothesis testing (details).
- Permutation test accounts for intra-correlations.
- Benjamini-Yekutieli correction for multiple testing.
- Interested in connected stable bimodules

Notation

$r(s,t)$: sample correlation of s, t

$r^2(A', B') \doteq \sum_{s \in A'} \sum_{t \in B'} r^2(s,t)$

Stable Bimodules



Stable bimodule (definition)

(A, B) is a *stable bimodule* if

$$A = \{s \in S \mid r^2(s, B) \text{ is significant}\}, \text{ and}$$

$$B = \{t \in T \mid r^2(A, t) \text{ is significant}\}.$$

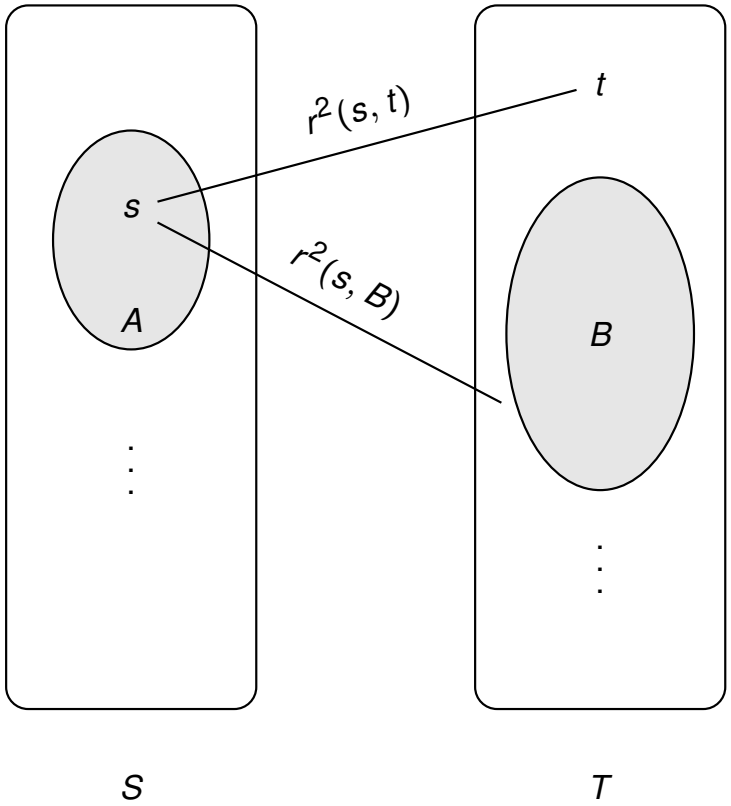
- Recursive definition like a community; made precise using hypothesis testing (details).
- Permutation test accounts for intra-correlations.
- Benjamini-Yekutieli correction for multiple testing.
- Interested in connected stable bimodules

Notation

$r(s,t)$: sample correlation of s, t

$r^2(A', B') \doteq \sum_{s \in A'} \sum_{t \in B'} r^2(s,t)$

Stable Bimodules



Stable bimodule (definition)

(A, B) is a *stable bimodule* if

$$A = \{s \in S \mid r^2(s, B) \text{ is significant}\}, \text{ and}$$

$$B = \{t \in T \mid r^2(A, t) \text{ is significant}\}.$$

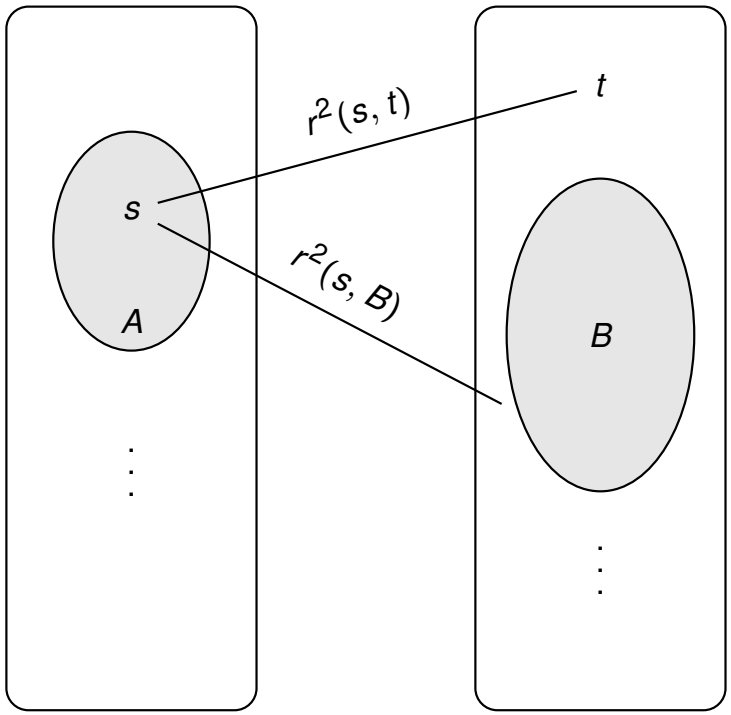
- Recursive definition like a community; made precise using hypothesis testing (details).
- Permutation test accounts for intra-correlations.
- Benjamini-Yekutieli correction for multiple testing.
- Interested in connected stable bimodules

Notation

$r(s,t)$: sample correlation of s, t

$r^2(A', B') \doteq \sum_{s \in A'} \sum_{t \in B'} r^2(s,t)$

Bimodule Search Procedure (BSP)



$$r^2(A,B) \doteq \sum_{s\in A}\sum_{t\in B} r^2(s,t)$$

Note, stability is just a fixed point condition:

$$A = \{s \in S \mid r^2(s,B) \text{ is significant}\} \doteq \Gamma_S(B)$$
$$B = \{t \in T \mid r^2(A,t) \text{ is significant}\} \doteq \Gamma_T(A).$$

Find stable bimodules by iterating

$$(A_k, B_k) = (\Gamma_S(B_k), \Gamma_T(A_{k-1})) \quad k = 1, 2, \ldots$$

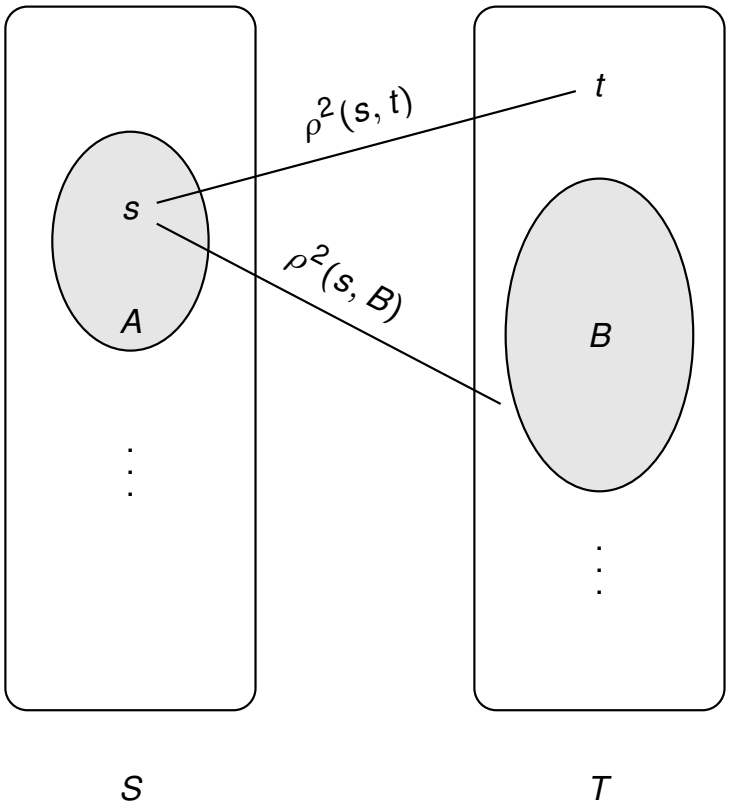
till sets don't change, for suitable $A_0 \subseteq S$.

Bimodule Search Procedure (BSP)

Starting from singletons $A_0 = \{s\} \in S$, iterate the definition till fixed point is reached (or sets cycle).

Covergence on real data | Example of an iterative search

Bimodule Search Procedure (BSP)



Population analysis ($n \to \infty$)
BSP iterations converge to connected components of the population correlation network.

Note, stability is just a fixed point condition:

$$A = \{s \in S \mid \rho^2(s, B) > 0\} \doteq \Gamma_S(B)$$
$$B = \{t \in T \mid \rho^2(A, t) > 0\} \doteq \Gamma_T(A).$$

Find stable bimodules by iterating

$$(A_k, B_k) = (\Gamma_S(B_k), \Gamma_T(A_{k-1})) \quad k = 1, 2, \ldots$$

till sets don't change, for suitable $A_0 \subseteq S$.

Bimodule Search Procedure (BSP)

Starting from singletons $A_0 = \{s\} \in S$, iterate the definition till fixed point is reached (or sets cycle).

Covergence on real data | Example of an iterative search

Data from GTEx project (v8)

from gtexportal.org

NIH funded GTEx project

A large collection of multi-tissue eQTL data from donors.

Individuals densely genotyped

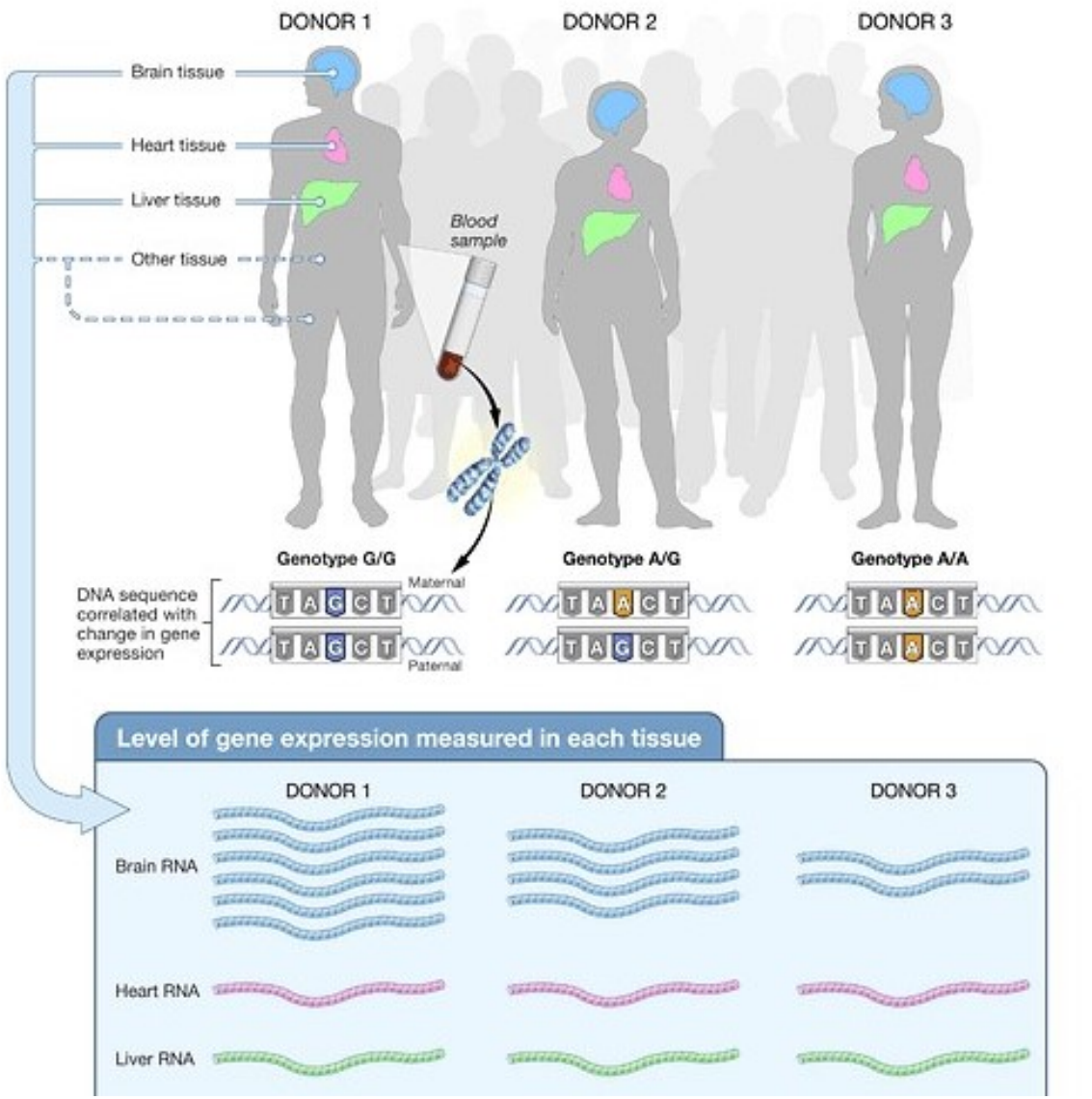
Measurements for 4.9 million SNPs encoded as $\{0, 1, 2\}$ (MAF).

Expression measured in multiple tissues

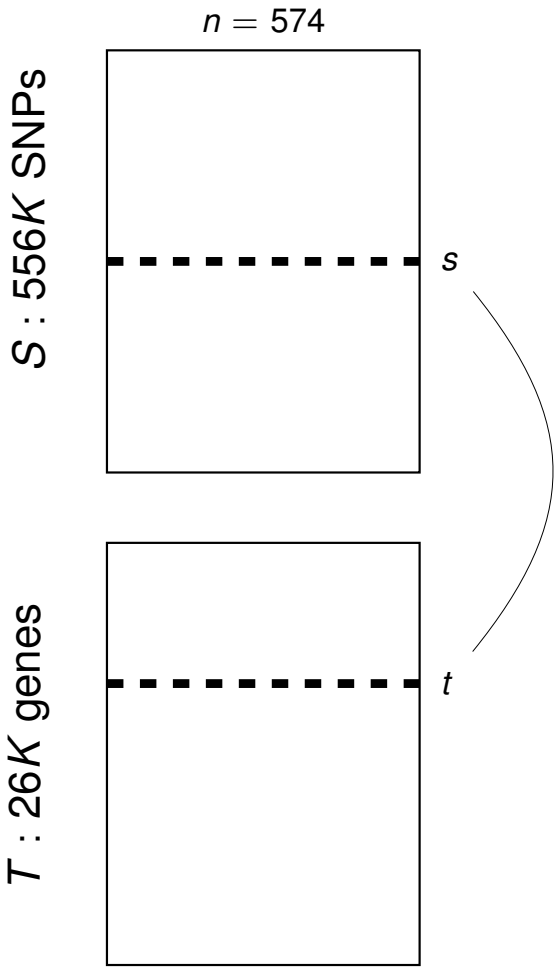
RNA sequencing used to measure expression of genes.

Normalization, quality control, and covariate correction performed.

Genomics glossary



eQTL analysis for Thyroid data



Thyroid expression data from $n = 574$ donors for

$T = \{26K \text{ genes}\}$ and

$S = \{556K \text{ representative SNPs}\}$ (after LD-pruning)

standard eQTL analysis

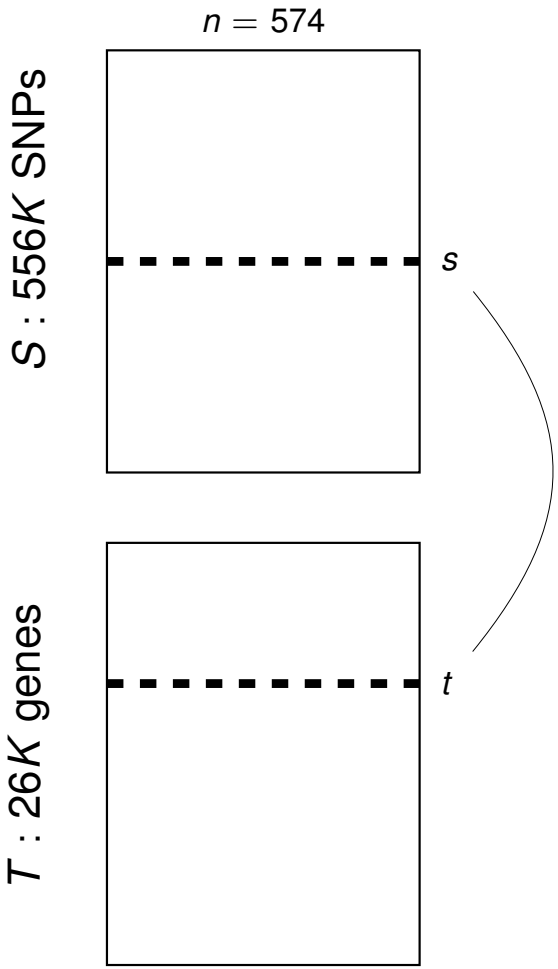
Find pairs $s \in S$ and $t \in T$ for which $r^2(s, t)$ is significant after correcting for multiple-testing (a statistical burden).

Finding SNP-gene bimodules (CONDOR)

Platig et al. (2016) find SNP-gene bimodules by community detection on a bipartite graph obtained from standard eQTL analysis.

They show that SNP-gene bimodules may have better functional interpretation than individual SNP-gene pairs.

eQTL analysis for Thyroid data



Thyroid expression data from $n = 574$ donors for

$T = \{26K \text{ genes}\}$ and

$S = \{556K \text{ representative SNPs}\}$ (after LD-pruning)

standard eQTL analysis

Find pairs $s \in S$ and $t \in T$ for which $r^2(s, t)$ is significant after correcting for multiple-testing (a statistical burden).

Finding SNP-gene bimodules (CONDOR)

Platig et al. (2016) find SNP-gene bimodules by community detection on a bipartite graph obtained from standard eQTL analysis.

They show that SNP-gene bimodules may have better functional interpretation than individual SNP-gene pairs.

Running BSP on GTEx Thyroid data

Highlights of results and validation

- BSP has a single free parameter $\alpha \in (0, 1)$ that was chosen using permutation to control a network-based false-discovery rate.
- Scatter plot BSP found 3305 bimodules in 4.7 hrs (20-core/2.4 GHz machine) of various sizes, having 1-1000 SNPs & 1-100 genes.
- Locations analysis Local and distal SNP-genes pairs in bimodules: most bimodules had at least one local SNP-gene pair, while larger bimodules had distal associations.
- Network analysis Connected SNP-gene networks underlying bimodules. Note: stable bimodule are defined in terms of aggregate associations, and all SNP-gene pairs in a bimodule do not have to be eQTLs.
- BSP vs. standard analysis BSP Bimodules vs. standard eQTL-analysis: most bimodules were connected under eQTLs, but new potential eQTLs were discovered by the remaining bimodules. Most of distal eQTLs, and half of local eQTLs were found by bimodules.
- GO analysis Gene ontology analysis : many bimodule were enriched for overlap with biological process related gene sets from the GO database, but the significant GO terms did not seem thyroid related.

Limit theorems for the Supermarket Model

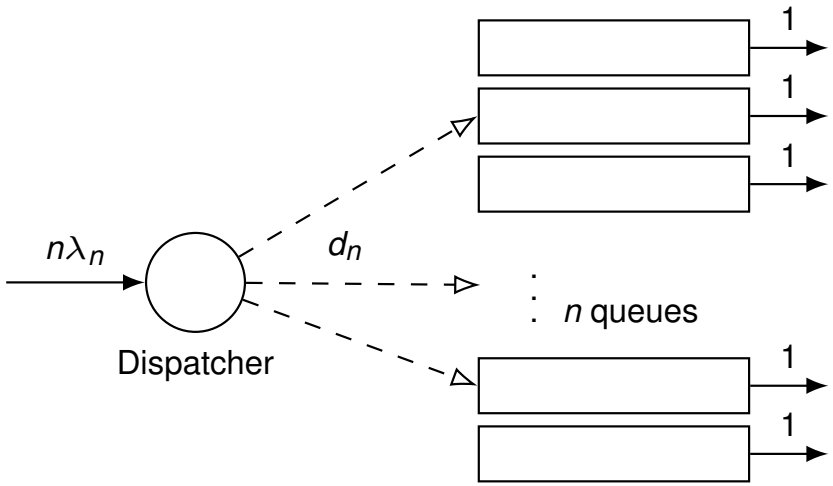


Figure: Supermarket model with parameters (n, d_n, λ_n).

Stochastic process for the Supermarket model

Convenient state descriptor

$G_{n,i}(t) =$ fraction of queues with length $\geq i$ at time t.

Observe:

$$1 = G_{n,0}(t) \geq G_{n,1}(t) \geq G_{n,2}(t) \ldots \geq 0$$

Hence $\boldsymbol{G}_n(t) \doteq (G_{n,i}(t))_{i \geq 1} \in \ell_1^{\downarrow}$, where

$$\ell_1^{\downarrow} \doteq \{(x_1, x_2, \ldots) \mid x_1 \geq x_2 \geq \ldots\} \cap \ell_1$$

Example: queues arranged in increasing order ($n = 10$).

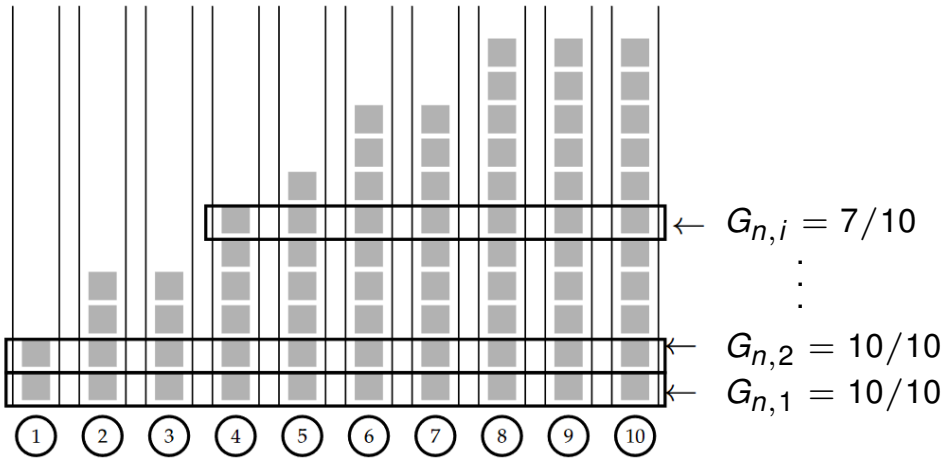


Image from Mukherjee et al. (2016)

Remarks

- $\boldsymbol{G}_n(t)$ is the empirical distribution of queue lengths at time t.
- Due to symmetry of queues, $\boldsymbol{G}_n(t)$ is an $\ell_1^{\downarrow}$ valued CTMC.

Stochastic process for the Supermarket model

Convenient state descriptor

$G_{n,i}(t) =$ fraction of queues with length $\geq i$ at time t.

Example: queues arranged in increasing order ($n = 10$).

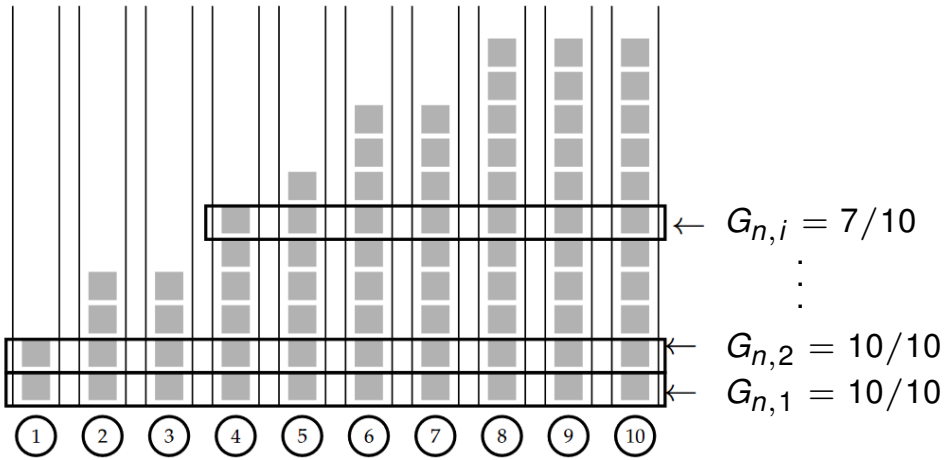


Image from Mukherjee et al. (2016)

Observe:

$$1 = G_{n,0}(t) \geq G_{n,1}(t) \geq G_{n,2}(t) \ldots \geq 0$$

Hence $\boldsymbol{G}_n(t) \doteq (G_{n,i}(t))_{i \geq 1} \in \ell_1^{\downarrow}$, where

$$\ell_1^{\downarrow} \doteq \{(x_1, x_2, \ldots) \mid x_1 \geq x_2 \geq \ldots\} \cap \ell_1$$

Remarks

- $\boldsymbol{G}_n(t)$ is the empirical distribution of queue lengths at time t.
- Due to symmetry of queues, $\boldsymbol{G}_n(t)$ is an $\ell_1^{\downarrow}$ valued CTMC.

Functional law of large numbers with growing choices

Previous work

First order behavior: want to show $\boldsymbol{G}_n \xrightarrow{P} \boldsymbol{g}$ in $D_{\ell_1^{\downarrow}}[0,\infty)$ as $n \to \infty$.

Mukherjee, Borst, Leeuwaarden, and Whiting (2016)

If $\lambda_n \to \lambda \in [0,1)$, $d_n \to \infty$, and suitable conditions hold at $t = 0$ as $n \to \infty$, then

- The sequence $\{\boldsymbol{G}_n\}_{n\in\mathbb{N}}$ is tight in the space of $\ell_1^{\downarrow}$ valued functions.
- If $\boldsymbol{G}_{n_k} \Rightarrow \boldsymbol{G}$ for some sub-sequence $\{n_k\}_{k\in\mathbb{N}}$, then $\boldsymbol{G}$ satisfies a certain differential equation.
- Universality: the differential equation does not depend on the rate at which $d_n \to \infty$.
- Interchange of limits: $\boldsymbol{G}_n(\infty) \Rightarrow \delta_{(\lambda,0,0,\ldots)}$ where $(\lambda, 0, 0, \ldots) \in \ell_1^{\downarrow}$ is the unique fixed point of the differential equation.

Uniqueness of the solution to the differential equation, and hence convergence of $\boldsymbol{G}_n$, was not ultimately shown.

Functional law of large numbers with growing choices

We complete the proof.

(Bhamidi, Budhiraja, D.) FLLN as $d_n \to \infty$ and $\lambda_n \to \lambda < \infty$

If $\boldsymbol{G}_n(0) \xrightarrow{P} (r_1, r_2, \ldots)$ in $\ell_1^{\downarrow}$, then $\boldsymbol{G}_n \xrightarrow{P} \boldsymbol{g}$ in $D_{\ell_1^{\downarrow}}[0,\infty)$ where $(\boldsymbol{g}, \boldsymbol{v}) \in C_{\ell_1^{\downarrow} \times \ell_\infty}[0,\infty)$ is the unique solution that solves the system

$$(g_i(t), v_i(t)) = (\Gamma_1, \hat{\Gamma}_1)\left(r_i - \int_0^{\cdot} (g_i(s) - g_{i+1}(s))ds + v_{i-1}(\cdot) \right)(t) \qquad \forall t > 0, i \in \mathbb{N}$$

and $v_0(t) = \lambda t$.

The limit satisfies reflected integral equations. Reflection comes from:

Skorokhod map (SM) for a function f with reflection from above at $\alpha \in \mathbb{R}$

$\Gamma_\alpha(f)(t) \doteq f(t) - \hat{\Gamma}_\alpha(f)(t) \leq \alpha$ is the reflected process, where
$\hat{\Gamma}_\alpha(f)(t) \doteq \sup_{s \in [0,t]} (f(s) - \alpha)^+$ is the minimal push.

Uniqueness follows from basic properties of SM.

FLLN with growing choices: proof idea

Representation in terms of time-change of Poisson process

There are independent unit-rate Poisson Process $\{N_{i,+}, N_{i,-}\}_{i\geq 1}$, so that for each $t > 0$

$$G_{n,i}(t) = G_{n,i}(0) + \frac{1}{n} N_{+,i}\left(n\lambda_n \int_0^t G_{n,i-1}^{d_n}(s) - G_{n,i}^{d_n}(s) ds\right) - \frac{1}{n} N_{-,i}\left(n \int_0^t G_{n,i}(s) - G_{n,i+1}(s) ds\right)$$

Semi-martingale decomposition using compensators of the point processes

$$G_{n,i}(t) = G_{n,i}(0) + \lambda_n \int_0^t (G_{n,i-1}^{d_n}(s) - G_{n,i}^{d_n}(s)) ds - \int_0^t (G_{n,i}(s) - G_{n,i+1}(s)) ds + M_{n,i}(t)$$

Martingale convergence and tightness of $\{\boldsymbol{G}_n\}_{n\geq 1}$

For any $T > 0$, $\sup_{t\in[0,T]} \|\boldsymbol{M}_n(t)\|_2 \xrightarrow{P} 0$ where $\boldsymbol{M}_n = (M_{n,1}, M_{n,2}, \ldots)$. Can then show that $\{\boldsymbol{G}_n\}_{n\in\mathbb{N}}$ is tight. Then by Skorokhod embedding assume $(\boldsymbol{G}_{n_k}, \boldsymbol{M}_{n_k}) \to (\boldsymbol{G}, \boldsymbol{0})$ a.s. uniformly.

Finally, show $\boldsymbol{G}$ satisfies the integral equation. Key step: Since $d_n \to \infty$,

$v_i(t) \doteq \lim_n \lambda_n \int_0^t G_{n,i}^{d_n}(s) ds$ satisfies $\int_0^t (1 - G_i(s)) dv_i(s) = 0$ and hence the SM emerges.

Functional central limit theorem with $d_n \gg \sqrt{n}\log n$

Previous work

Assume $\sqrt{n}(1-\lambda_n) \to \beta > 0$ (Halfin-Whitt regime)

Let $\boldsymbol{Y}_n \doteq \sqrt{n}(\boldsymbol{G}_n - \boldsymbol{e}_1)$ (hence we expect $G_{n,1} = 1 - o(1)$ and $G_{n,2} = o(1)$)

Assume $\boldsymbol{Y}_n(0)$ converges in probability.

Eschenfeldt and Gamarnik (2015)

If $d_n = n$ then as $n \to \infty$

$$\boldsymbol{Y}_n \Rightarrow (Y_1, Y_2, 0, \ldots) \qquad \text{in a suitable function space.}$$

where (Y_1, Y_2) is a 2D reflected diffusion process driven by a 1D Brownian motion.

Limiting system

Mukherjee, Borst, Leeuwaarden, and Whiting (2016)

- The above result continues to hold if $d_n \gg \sqrt{n}\log n$.
- $\{\boldsymbol{Y}_n\}_{n\geq 1}$ is not tight if $d_n \ll \sqrt{n}\log n$.

Center around Near Equilibrium point for $d_n \leq \sqrt{n} \log n$

Recall, semimartingale representation for our system:

$$\boldsymbol{G}_n(t) = \boldsymbol{G}_n(0) + \int_0^t [\boldsymbol{a}_n(\boldsymbol{G}_n(s)) - \boldsymbol{b}(\boldsymbol{G}_n(s))]ds + \boldsymbol{M}_n(t)$$

Near Equilibrium (NE) point $\boldsymbol{\mu}_n \in \ell_1^{\downarrow}$ is the unique solution to

$$\boldsymbol{a}_n(\boldsymbol{\mu}_n) = \boldsymbol{b}(\boldsymbol{\mu}_n)$$

Remarks

- Since the drift vanishes in the first display and $\boldsymbol{M}_n \xrightarrow{P} 0$, if $\boldsymbol{G}_n(0) \approx \boldsymbol{\mu}_n$ we expect $\boldsymbol{G}_n(t) \approx \boldsymbol{\mu}_n$ over compact times t.
- $\boldsymbol{\mu}_n \in \ell_1^{\downarrow}$ is the state where inflow rate equals the outflow rate for each coordinate. E.g. when sampling with replacement $\boldsymbol{\mu}_n = (\lambda_n, \lambda_n^{d_n}, \lambda_n^{d_n+d_n^2}, \lambda_n^{d_n+d_n^2+d_n^3}, \ldots) \in \ell_1^{\downarrow}$.
- Convergence will be shown for the process $\boldsymbol{Z}_n \doteq \sqrt{n}(\boldsymbol{G}_n - \boldsymbol{\mu}_n)$ for $1 \ll d_n \leq n$.

FCLT for $d_n \gg \sqrt{n}$ centered around NE points

(Bhamidi, Budhiraja, D.) FCLT as $d_n \gg \sqrt{n}$ and $1 - \lambda_n = \frac{\log d_n}{d_n} + \frac{\alpha_n}{\sqrt{n}}$

If $\alpha_n \to \alpha \in [0, \infty]$, $\boldsymbol{Z}_n(0) \xrightarrow{P} (z_1, z_2, 0, 0, \ldots) \in \ell_2$ and $z_1 \leq \alpha$ as $n \to \infty$, then $\boldsymbol{Z}_n \Rightarrow (Z_1, Z_2, 0 \ldots)$ in $D_{\ell_2}[0, \infty)$ where (Z_1, Z_2, η) the reflected diffusion process

$$(Z_1(t), \eta(t)) = (\Gamma_\alpha, \hat{\Gamma}_\alpha)\left(z_1 - \int_0^{\cdot} (Z_1(s) - Z_2(s))ds + \sqrt{2}B(\cdot)\right)(t)$$

$$Z_2(t) = z_2 + \eta(t) - \int_0^t Z_2(s)ds,$$

where B is a 1D standard Brownian motion.

Remarks

- Allows for $(\lambda_n, d_n) = (1 - n^{-a}, n^{b_n})$ if $a \in [1/3, 1)$ and $b_n \in [a \vee 0.5 + \frac{\ln \ln n}{\ln n}, 1]$.
- Limit is similar to Eschenfeldt & Gamarnik (2015), but reflection at α and a missing drift term.
- Easy to show $\boldsymbol{Z}_n - \boldsymbol{Y}_n \to -\alpha \boldsymbol{e}_1$ when $d_n \gg \sqrt{n} \log n$ and $\sqrt{n}(1 - \lambda_n) \to \alpha$. Hence the results in Mukherjee et al. (2016), Eschenfeldt & Gamarnik (2015) follow as a special case.

FCLT for $\frac{d_n}{\sqrt{n}} \to c \in (0, \infty)$ around NE points

(Bhamidi, Budhiraja, and D.) FCLT as $d_n \sim c\sqrt{n}$ and $1 - \lambda_n = \frac{\log d_n}{d_n} + \frac{\alpha_n}{\sqrt{n}}$

If $\alpha_n \to \alpha \in (-\infty, \infty]$ and $\boldsymbol{Z}_n(0) \xrightarrow{P} (z_1, z_2, 0, \ldots)$ as $n \to \infty$, then $\boldsymbol{Z}_n \Rightarrow (Z_1, Z_2, 0 \ldots) \in D_{\ell_2}[0, \infty)$, where (Z_1, Z_2) is the unique pathwise solution to

$$Z_1(t) = z_1 - \int_0^t (Z_1(s) - Z_2(s))ds - (ce^{c\alpha})^{-1} \int_0^t \left(e^{cZ_1(s)} - 1\right) ds + \sqrt{2}B(t)$$

$$Z_2(t) = z_2 - \int_0^t Z_2(s)ds + (ce^{c\alpha})^{-1} \int_0^t \left(e^{cZ_1(s)} - 1\right) ds$$

for a 1D standard Brownian motion B.

Remarks

- For $\alpha < \infty$, R.H.S. is not a Lipschitz function of (Z_1, Z_2). But since the exponential term opposes the growth of the system, the solution is well defined for a.e. sample path.
- Applies with $\alpha = \infty$ to the case $\lambda_n = 1 - n^{-a}$ for fixed $a \in (1/4, 1/2)$ (sub-Halfin-Whitt).

FCLT for $1 \ll d_n \ll \sqrt{n}$ around NE points

(Bhamidi, Budhiraja, D.) FCLT as $1 \ll d_n^{k+1} \ll n$, $1 - \lambda_n = \frac{\log d_n - \xi_n}{d_n^k}$, and $k \in \mathbb{N}$

If $e^{\xi_n} \to \alpha \in [0, \infty)$, $\boldsymbol{Z}_n(0) \xrightarrow{P} (z_1, \ldots z_{k+1}, 0, 0, \ldots)$, then $(Z_{n,1}, \ldots, Z_{n,k-1}) \Rightarrow \mathbf{0}$ in $D_{\mathbb{R}^{k-1}}(0, \infty)$ and $(\sum_{j=1}^{k} Z_{n,i}, Z_{n,k+1}, Z_{N,k+2}, \ldots) \Rightarrow (X_1, X_2, 0, \ldots)$ in $D_{\ell_2}[0, \infty)$, where (X_1, X_2) is the diffusion

$$X_1(t) = \sum_{i=1}^{k} z_i - (\alpha + \mathbb{I}_{\{k=1\}}) \int_0^t X_1(s)ds + \int_0^t X_2(s)ds + \sqrt{2}B(t)$$

$$X_2(t) = z_{k+1} + \alpha \int_0^t X_1(s)ds - \int_0^t X_2(s)ds.$$

where B is a 1D standard Brownian motion.

Remark

- The NE point converges to $\mu_n \to \sum_{i=1}^{k} \boldsymbol{e}_i = (1, \ldots, 1, 0, \ldots)$. Previous theorems had $k = 1$.
- Most queues will have length k. Brightwell et al. (2018) show similar result for the equilibrium.
- Applies to $\lambda_n = 1 - n^{-a}$ for $a \in (0, 1)$, $k \geq a/(1-a)$ and $d_n = (n^a \log n)^{1/k}$.
- Limit of the first $k - 1$ coordinates "instantly" become 0.

FCLT for $1 \ll d_n \ll \sqrt{n}$ around NE points

(Bhamidi, Budhiraja, D.) FCLT as $1 \ll d_n^{k+1} \ll n$, $1 - \lambda_n = \frac{\log d_n - \xi_n}{d_n^k}$, and $k \in \mathbb{N}$

If $e^{\xi_n} \to \alpha \in [0, \infty)$, $\boldsymbol{Z}_n(0) \xrightarrow{P} (z_1, \ldots z_{k+1}, 0, 0, \ldots)$, then $(Z_{n,1}, \ldots, Z_{n,k-1}) \Rightarrow \mathbf{0}$ in $D_{\mathbb{R}^{k-1}}(0, \infty)$ and $(\sum_{j=1}^{k} Z_{n,i}, Z_{n,k+1}, Z_{N,k+2}, \ldots) \Rightarrow (X_1, X_2, 0, \ldots)$ in $D_{\ell_2}[0, \infty)$, where (X_1, X_2) is the diffusion

$$X_1(t) = \sum_{i=1}^{k} z_i - (\alpha + \mathbb{I}_{\{k=1\}}) \int_0^t X_1(s)ds + \int_0^t X_2(s)ds + \sqrt{2}B(t)$$

$$X_2(t) = z_{k+1} + \alpha \int_0^t X_1(s)ds - \int_0^t X_2(s)ds.$$

where B is a 1D standard Brownian motion.

Remark

- The NE point converges to $\boldsymbol{\mu}_n \to \sum_{i=1}^{k} \boldsymbol{e}_i = (1, \ldots, 1, 0, \ldots)$. Previous theorems had $k = 1$.
- Most queues will have length k. Brightwell et al. (2018) show similar result for the equilibrium.
- Applies to $\lambda_n = 1 - n^{-a}$ for $a \in (0, 1)$, $k \geq a/(1-a)$ and $d_n = (n^a \log n)^{1/k}$.
- Limit of the first $k - 1$ coordinates "instantly" become 0.

Conclusion

Finding stable bimodules in multi-view data

- Bimodule: a group of features in bi-view data with significant aggregate cross-correlation, and a community in the cross-correlation network.
- Bimodule Search Procedure. Finds *stable and connected* bimodules – a fixed point condition based on hypothesis tests. Parallel R implementation.
- Application to eQTL analysis. SNP-gene bimodules may provide more insights than traditional pairwise analysis.
- Future directions: Theoretical false discovery guarantees for the iterative search and/or stable bimodules. Extensions to multi-view data and other metrics like co-occurance.

Limit theorems for the Supermarket model with growing choices

- Supermarket model: Model for load balancing in large data centers, based on the randomized 'Power of d choices' routing.
- Limit theorems in heavy traffic to understand model behavior when the number of choices d increases with system size.
- Future directions: Show interchange of limits and convergence of stationary distributions. Prove similar limit theorems for related models like the Supermarket model with memory.

Thank you!

BSP manuscript `https://arxiv.org/abs/2009.05079`

BSP R Package `https://github.com/miheerdew/cbce`.

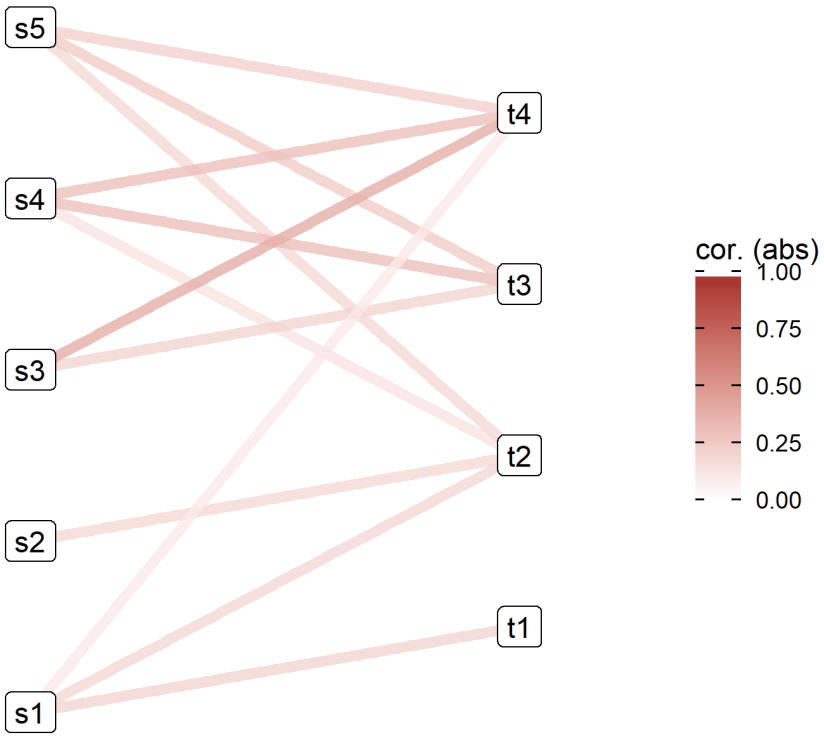
Limit theorems manuscript `https://arxiv.org/abs/2006.03621`

Supporting Grants

- NIH R01 HG009125-01
- NSF DMS-1613072
- SAMSI

Appendix

Example of a BSP iteration

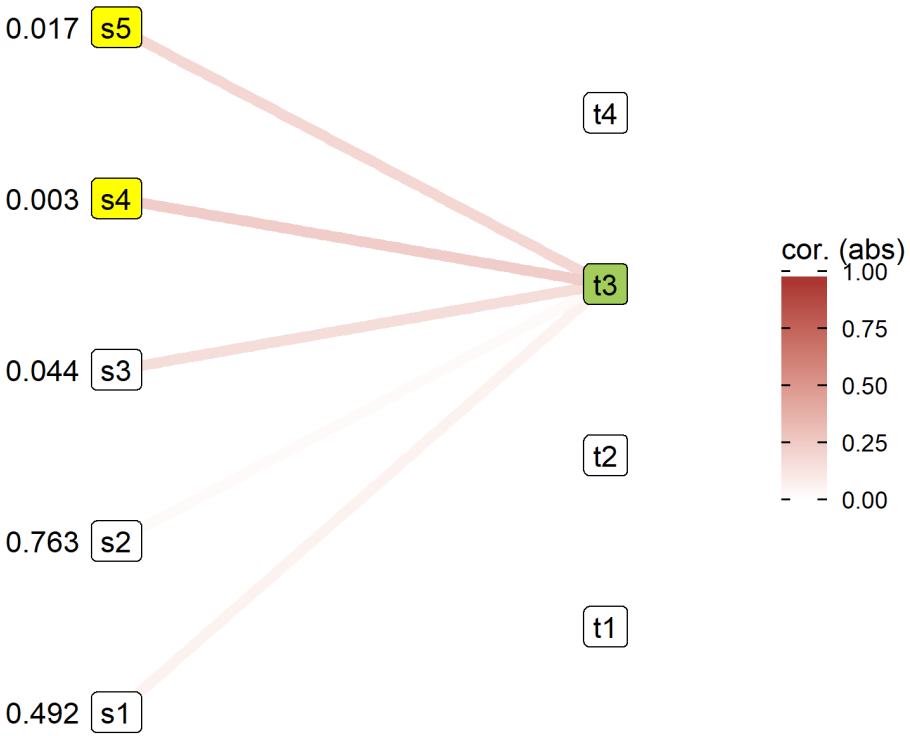


1. $B_0 = \{G_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{G_3, G_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{G_3, G_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Example of a BSP iteration

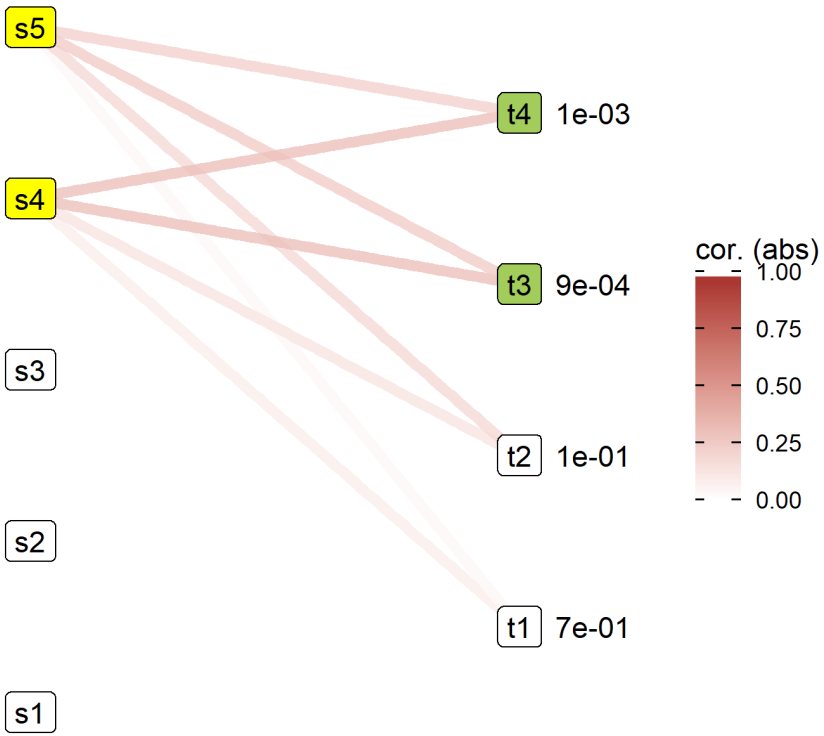


1. $B_0 = \{G_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{G_3, G_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{G_3, G_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Example of a BSP iteration

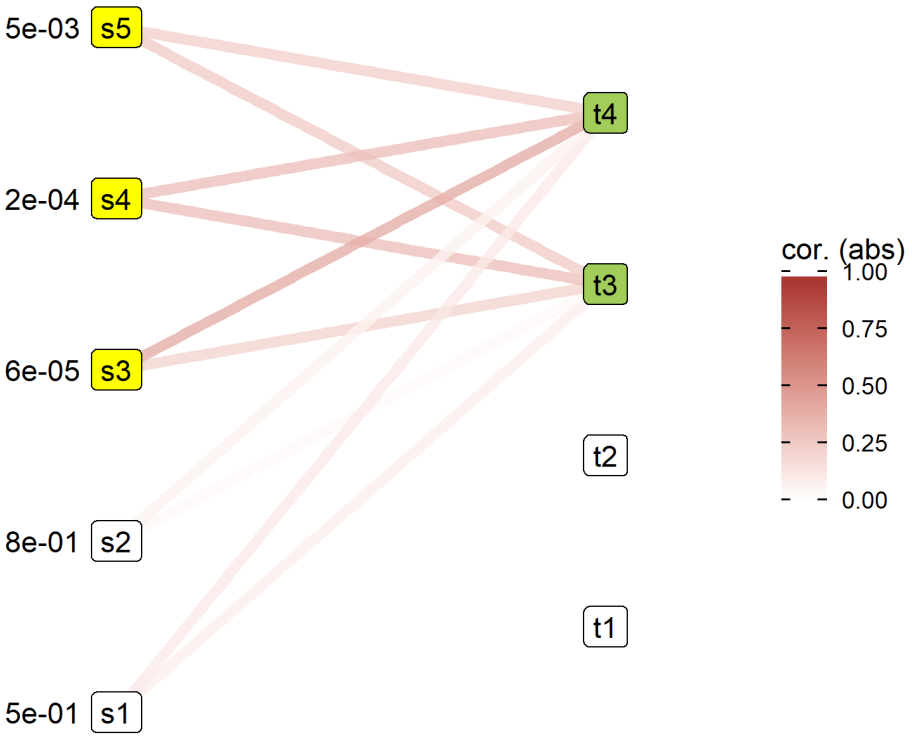


1. $B_0 = \{G_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{G_3, G_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{G_3, G_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Example of a BSP iteration

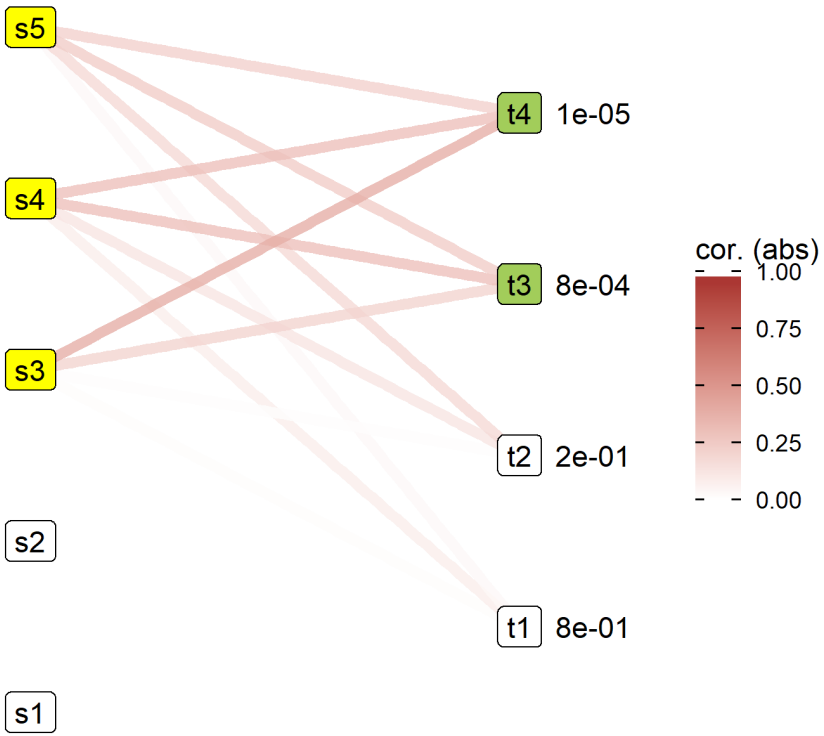


1. $B_0 = \{G_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{G_3, G_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{G_3, G_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Example of a BSP iteration

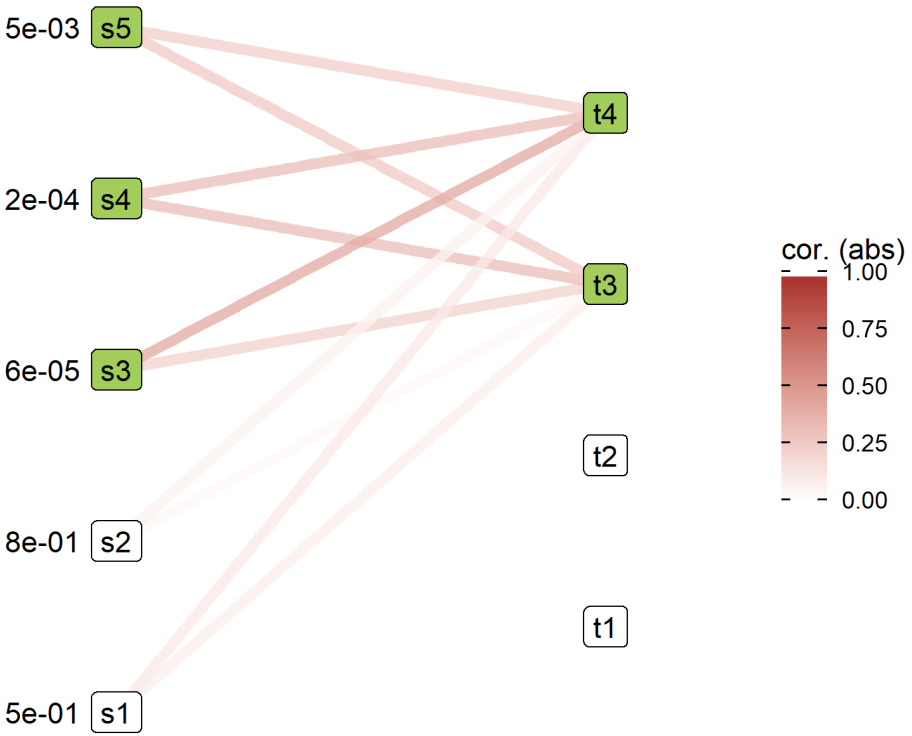


1. $B_0 = \{G_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{G_3, G_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{G_3, G_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

Example of a BSP iteration



1. $B_0 = \{G_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{G_3, G_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{G_3, G_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

$(A_1, B_1) = (A_2, B_2)$

Stable bimodule found.

BSP implementation details

- Start from all singletons $\{s\}$ in SNPs and $\{g\}$ in Genes, to find a bimodule list $\mathcal{B}$ (possibly empty).
- Bimodules often repeat in $\mathcal{B}$, so we filter duplicates:
  1. Determine effective number:

$$N_{eff} = \sum_{(A,B)\in\mathcal{B}} \sum_{a\in A, b\in B} (|A||B|N(a,b))^{-1}$$

  2. Hierarchical-cluster elements of $\mathcal{B}$ based on Jaccard distances.
  3. Select a height to cut the dendrogram so that N_{eff} clusters are made.
- R package with fast implementation : https://github.com/miheerdew/cbce.

Analysis of genomic locations of bimodules

Recall BSP does not use genomic locations of SNPs and Genes. Nevertheless

Proximity of SNPs and genes within the bimodule.

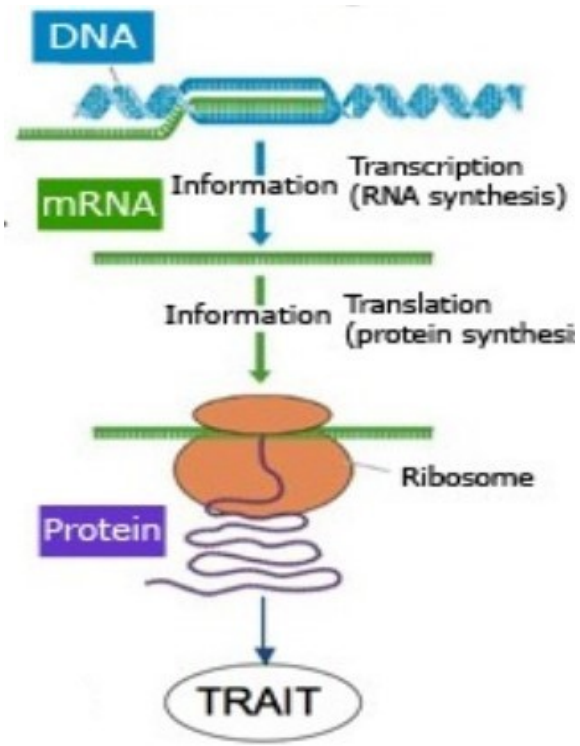
- Almost all (99.3%) bimodules have at least one local SNP-gene pair.
- In addition, almost half of the larger bimodules found gene and SNPs that had distal effects.

Chromosomal locations of SNPs and genes from bimodules.

- Bimodule SNPs and Genes distributed across all 23 chromosomes.
- Most small bimodules (95%) were restricted to single chromosome.
- Nearly half of the larger bimodules spanned 2-11 chromosomes each.

Concepts from genomics (simplified version)

genome.gov/genetics-glossary



Gene expression

Process used by cells to assemble protein molecules based on a gene.

Gene A region of the genome that encodes for a protein; ∼30K genes identified in humans.

Single nucleotide polymorphism (SNP)

A location on the genome that has a nucleotide variation within the population.

Genetic basis of gene expression

Millions of SNPs are identified in humans. Which ones influence traits?

Expression quantitative trait loci (eQTL)

A genomic region (e.g. SNP) that influences the expression level of one or more genes.

Obtaining networks from bimodules

A SNP-gene bimodule (A, B) has aggregate correlation between A and B.

But which edges $(s, t) \in A \times B$ are significant?

Threshold at $\tau \in (0, 1)$: $E_\tau(A, B) = \{(s, t) \mid r^2(s, t) \geq \tau^2,\ s \in A, t \in B\}$

How to choose τ?

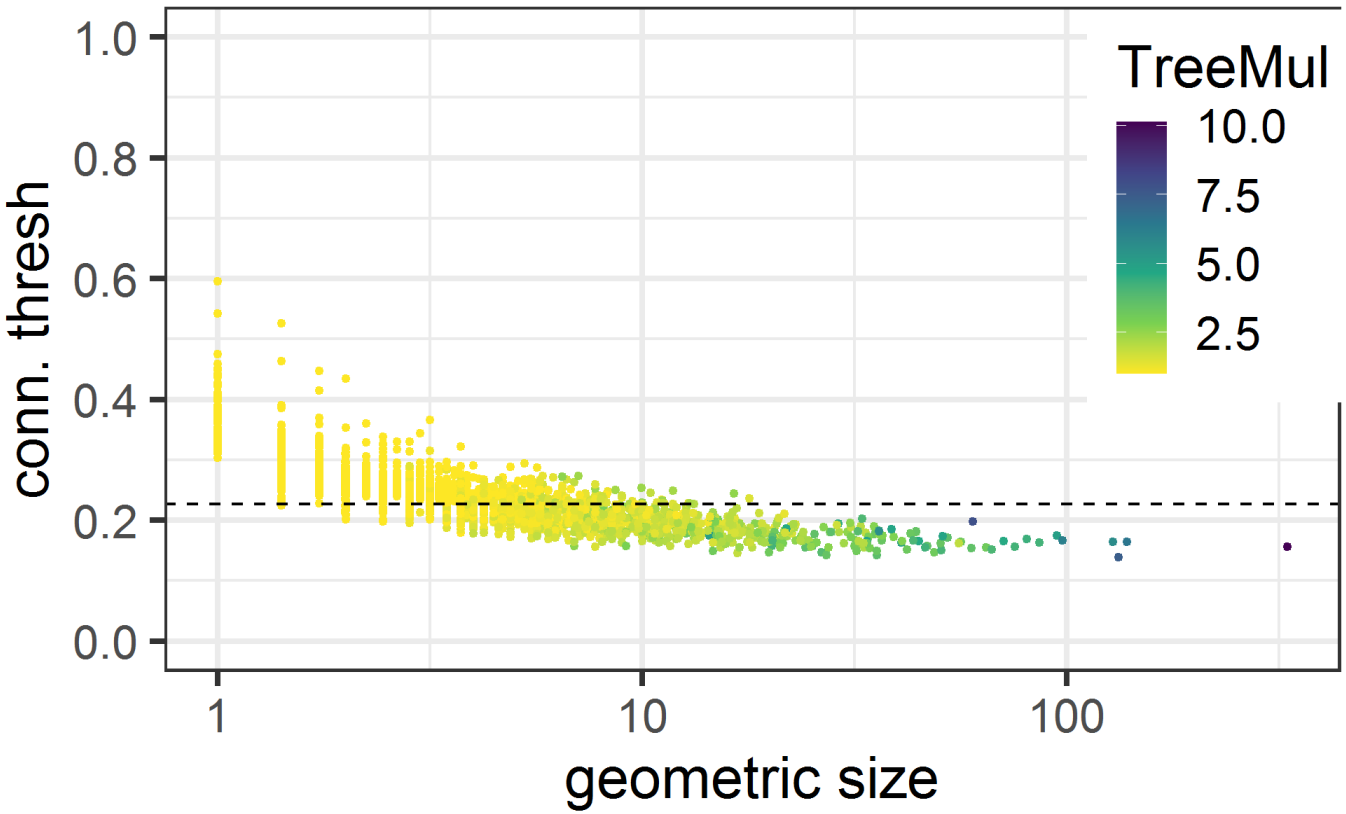
Conservative estimate of strongest edges

Since a bimodule must be connected, choose the largest $\tau^* \in (0, 1)$ so that $(A \sqcup B, E_{\tau^*}(A, B))$ is a connected graph.

$E_{\tau^*}(A, B)$ are called *essential-edges* of the bimodule.

Thyroid network statistics

Network statistics from BSP bimodules on GTEx data



Smaller bimods are connected mainly by strong local associations (large τ^*). E_{τ^*} is tree-like.

Larger bimods are connected by strong local + weak distal associations (small τ^*). E_{τ^*} has upto 10x more edges than a tree.

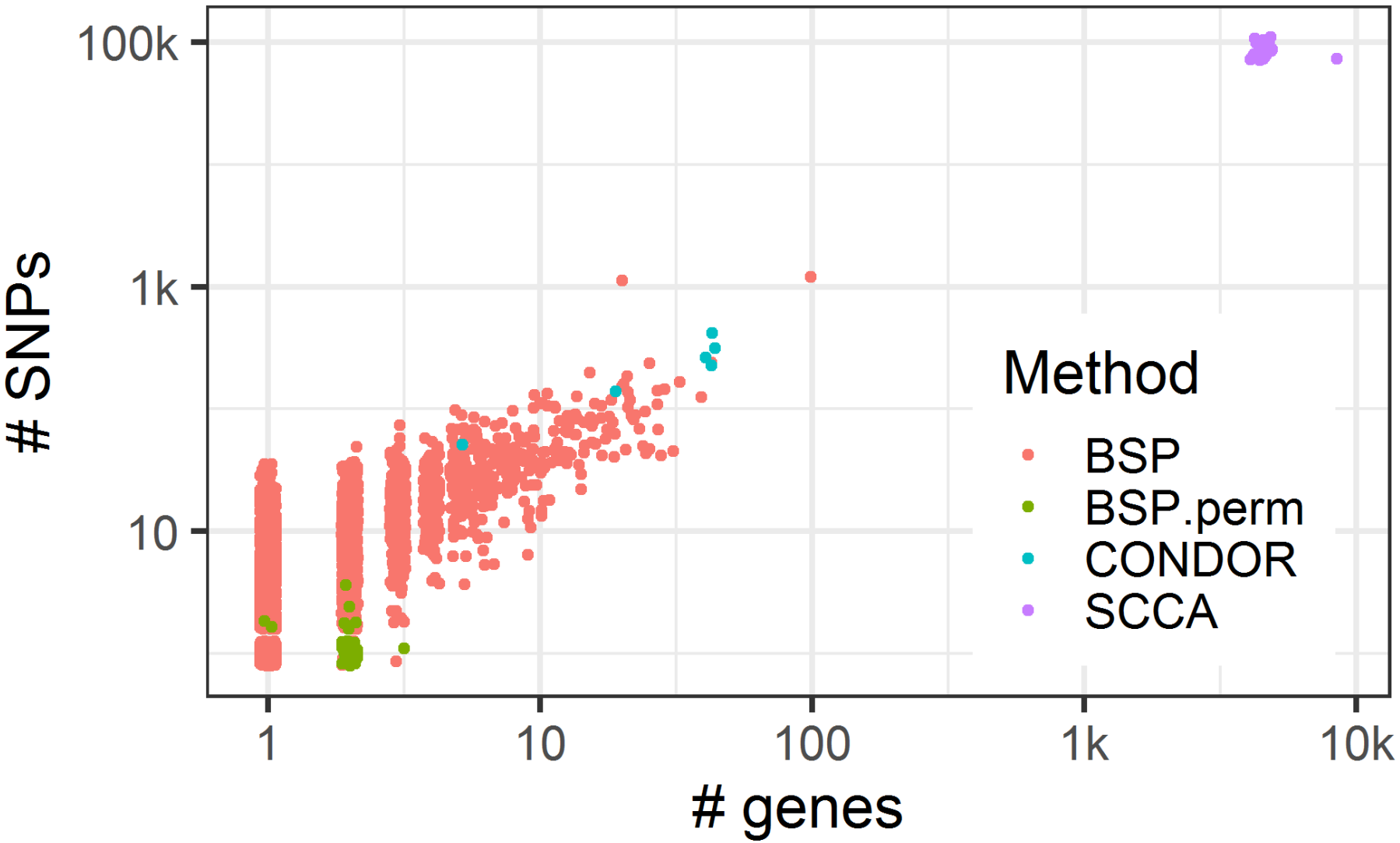
Enrichment of known gene sets in bimodules

The GO database (http://geneontology.org/) contains collection of gene sets known to be associated with biological functions.

- Consider our 145 bimodules that have 7 or more genes.
- We used Fisher's test to assess overlap of gene sets from these bimodules with GO sets.
- Gene sets from 18 bimodules had significant overlap with gene sets associated to known biological processes.
- But the associated function did not seem thyroid relevant.

Repeating above process with randomly chosen gene sets of the similar sizes did not detect significant association.

Sizes of bimodules discovered by various methods



BSP Thyroid search details

Search details

- 304K attempted searches.
- Majority (277K) give empty set in the first iteration.
- Few (20) did not terminate within 20 iterations.
- Remaining reached a fixed point in 20 iterations.
- 92.3% of these fixed points contained the seed singleton.

Assessing significance using hypothesis testing

How to quantify Γ_T?

$$\Gamma_T(A) \doteq \{t \in T \mid r^2(A,t) \text{ is significant}\}.$$

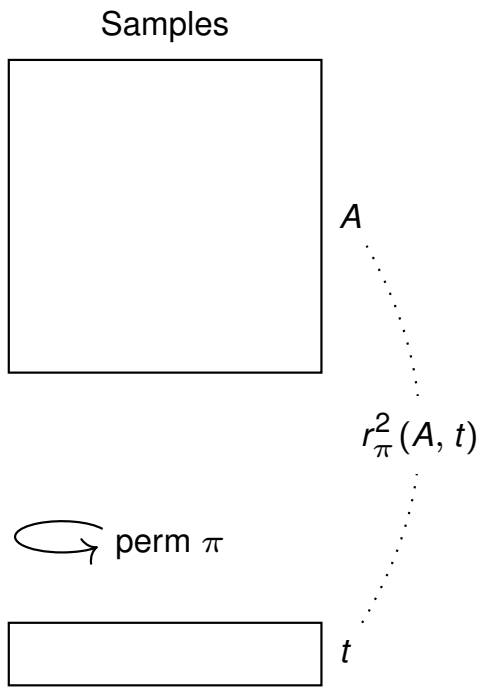
Steps

1. $\forall t \in T$ obtain p-value $p(A,t)$ from $r^2(A,t)$.
2. reject p-values using multiple-testing correction γ_α

$$\Gamma_T(A) = \{t \in T \mid p(A,t) \leq \gamma_\alpha\}$$

at some level $\alpha \in (0,1)$.

$p(A,t)$ conditional on intra-correlations in A



Permutation p-value

$$\mathbb{P}_\pi\left(r^2_\pi(A,t) \geq r^2_{obs}(A,t)\right)$$

Fast computation + other details

Hypothesis testing implementation details

Permutation p-values Permute sample labels of t using π. Define p-value

$$p_A(t) \doteq \mathbb{P}_\pi \left(r_\pi^2(A,t) \geq r^2(A,t) \right),$$

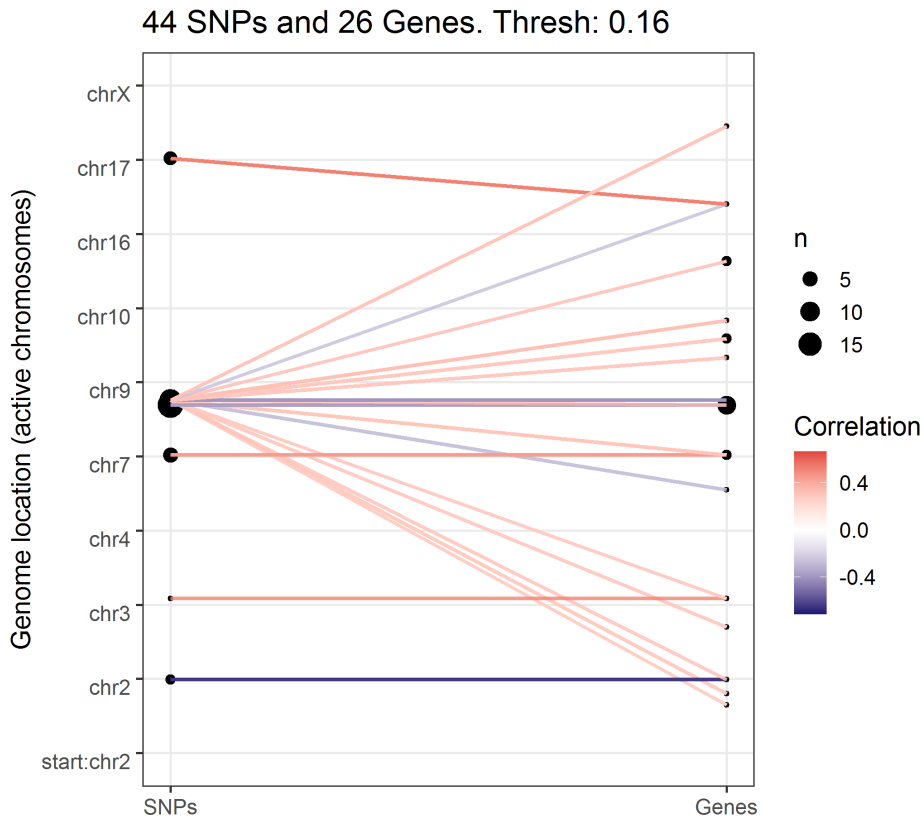
which conditions on correlations in A.

Multiple testing correction The adaptive threshold γ_α chosen from [Benjamini and Yekutieli, 2001] controls FDR at α.

Monte-Carlo estimation too slow. We fit a shifted gamma distribution to $T = r_\pi^2(A,t)$ based on top 3 moments. Moments of T are analytical approximated [Zhou, Gallins and Wright, 2019].

Essential-edge networks in GTEx thyroid data

examples from two bimodules



Comparing bimodules to standard eQTL analysis

Standard eQTL analysis performed using MatrixEQTL ($\alpha = 0.05$).

Bimodules find most standard eQTLs

84% of eQTLs from trans-analysis, and 51% of eQTLs from cis-analysis. But note

- bimodules find SNP-gene networks not just pairs, and
- cis-analysis improves power by restricting to local pairs.

New potential eQTLs from bimodules

224/358 large bimodules are not connected under edges from standard cis+trans analysis.

Essential-edges from bimodules reveal 300 local and 8.8k distal SNP-gene pairs that

- are not detected by standard analysis,
- but show significance at the network level.

FCLT for $d_n = n$ as $\lambda_n = 1 - \frac{\beta}{\sqrt{n}}$

Theorem (Eschenfeldt and Gamarnik 2018)

For some $\beta > 0$, let $d_n = n$ and $\lambda_n = 1 - \frac{\beta}{\sqrt{n}}$. If $\mathbf{Y}_n(0) \Rightarrow (y_1, y_2, 0, 0 \ldots)$ as $n \to \infty$, then $\mathbf{Y}_n \Rightarrow (Y_1, Y_2, 0, 0 \ldots)$ as processes as $n \to \infty$, where B is a SBM and for each $t > 0$

$$Y_1(t) = y_1 - \beta t - \int_0^t (Y_1(s) - Y_2(s))ds + \sqrt{2}B(t) - \eta(t) \tag{1}$$

$$Y_2(t) = y_2 + \eta(t) - \int_0^t Y_2(s)ds,$$

$Y_1(t) \leq 0$, and $\eta(t) = \int_0^t \mathbb{I}_{\{Y_1(s)=0\}} |d\eta|(s)$ is the smallest non-decreasing process that keeps (1) r.h.s. ≤ 0.

1. Limit is a constrained 2D system driven by 1D BM.
2. Since $G_{N,i}(t)$: fraction of queues with length $\geq i$ at time t
   by theorem $G_{N,1} = 1 - O_p(1/\sqrt{n})$ and $G_{n,2} = O_p(1/\sqrt{n})$.
3. Theorem is used to show that average waiting time is $O(1/\sqrt{n})$, same order as that of a $M/M/n$ system.