Guided state-space exploration of closed-loop control systems & Groupwise cross-correlation mining in bi-view data

Miheer Dewaskar

UNC Chapel Hill & Duke University

26th July 2022
SysConTalks @ IIT Bombay

Overview of projects

Guided state space exploration of closed-loop control systems

Joint work with **Manish Goyal and Parasara Sridhar Duggirala**.

Groupwise cross-correlation mining in bi-view data

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

Outline

1. Guided State-Space Exploration in Closed Loop Systems
   - Introduction to Cyber-Physical Systems (CPS)
   - Verifying safety of CPS
   - State space exploration using sensitivity
   - NExG: state space exploration by learning sensitivity direction

2. Groupwise cross-correlation mining in bi-view data
   - Bi-view data and Bimodules
   - Bimodules as (almost) communities in the cross-correlation network
   - Stable Bimodules and the Bimodule Search Procedure
   - eQTL analysis using GTEx data

Outline

1. Guided State-Space Exploration in Closed Loop Systems
    - Introduction to Cyber-Physical Systems (CPS)
    - Verifying safety of CPS
    - State space exploration using sensitivity
    - NExG: state space exploration by learning sensitivity direction

2. Groupwise cross-correlation mining in bi-view data
    - Bi-view data and Bimodules
    - Bimodules as (almost) communities in the cross-correlation network
    - Stable Bimodules and the Bimodule Search Procedure
    - eQTL analysis using GTEx data

Cyber-Physical Systems

Cyber-Physical Systems (CPS) are integrations of computation, networking, and physical processes. Embedded computers and networks monitor and control the physical processes, with feedback loops where physical processes affect computations and vice versa. [Derler et al., 2011]

Examples

Airplanes, medical monitoring, unmanned aerial vehicles, smart grid, and autonomous cars.

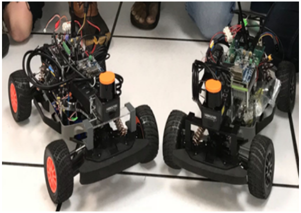


Figure: https://f1tenth.org/

Technical Challenges in modelling Cyber Physical Systems (CPS)

From https://ptolemy.berkeley.edu/projects/cps//

> *The key technical challenge [in modelling CPS] is to conjoin abstractions that have evolved over centuries for modeling physical processes (differential equations, stochastic processes, etc.) with abstractions that have evolved over decades in computer science [...]. The former abstractions focus on dynamics (evolution of system state over time), whereas the latter focus on processes of transforming data.*

Requires tools at the intersection of Computer Science, Statistics and Dynamical Systems, etc.; but is not simply a repackaging of old tools:

- 2013 lecture by Prof. Edward A. Lee, titled "Cyber-Physical Systems: A Fundamental Intellectual Challenge"

Outline

1. Guided State-Space Exploration in Closed Loop Systems
   - Introduction to Cyber-Physical Systems (CPS)
   - Verifying safety of CPS
   - State space exploration using sensitivity
   - NExG: state space exploration by learning sensitivity direction
2. Groupwise cross-correlation mining in bi-view data
   - Bi-view data and Bimodules
   - Bimodules as (almost) communities in the cross-correlation network
   - Stable Bimodules and the Bimodule Search Procedure
   - eQTL analysis using GTEx data

Our focus: Verifying safety of Cyber-Physical Systems (CPS)

It is important to ensure that the CPS that we deploy in the real world operate safely, and do not malfunction in unexpected or adversarial scenarios.

Verifying this for CPS is challenging because of the increasing complexity of the underlying systems as well as complex control algorithms (e.g. Neural Network controllers).

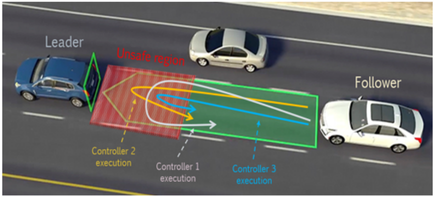
Examples of mishaps caused by faults in controlling software:

Avionics	Ariane 5, 1996 – software bug in the rocket's Inertial Reference System.
Industrial	Three Mile Island, 1979 – A malfunction and operator error resulted in inadequate cooling water circulation to the reactor core causing it to overheat and suffer a partial meltdown.
Medical	THERAC-25 1986, a machine for radiation therapy, caused accidents.

Can we design tools that automatically search for 'unsafe' executions (if they exist)?

Example of a linear closed-loop control system

Adaptive Cruise Control*: Car takes information from other cars as input and updates its speed accordingly



$$\dot{s} = (v_f - v)$$
$$\dot{v} = a$$
$$\dot{a} = g_1 a + g_2(v - v_f) + g_3(s - (v + 10))$$

v_f : leading car's velocity
v : follower's speed
a: follower's acceleration
s: distance

Initial Configuration	Controller 1	Controller 2
$s \in [2.1, 5]$	$g_1 = -3$	$g_1 = -1$
$v \in [18, 22]$	$g_2 = -3$	$g_2 = -3$
$v_f = 20$	$g_3 = 1$	$g_3 = 1$
$a \in [-1, 1]$		

*A. Tiwari, Approximate reachability for linear systems. HSCC, 2003

Mathematical formulation

System dynamics

The system state $x(t) \in \mathbb{D} \subseteq \mathbb{R}^n$ at time t evolves as

$$\dot{x} = f(x, u) \qquad (1)$$

where $u \in \mathbb{R}^m$ is the input to the system.

Closed-loop assumption

Suppose that we use a feedback-function $u = g(x)$ that is regulated by the system output. For every $x_0 \in \mathbb{D}$

$$x(t) = \xi(x_0, t) \quad t \in \mathbb{R}$$

denotes the unique trajectory satisfying (1) with $u = g(x)$ and $x(0) = x_0$.

Fix

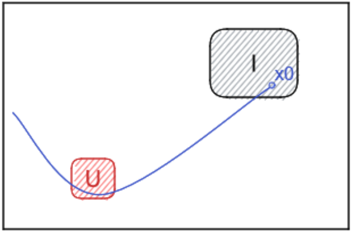
$I \subseteq \mathbb{D}$: the set of initial system states

$U \subseteq \mathbb{D}$: the collection of "unsafe" states

$T > 0$: a terminal time

Finding unsafe trajectories

Does there exist $x_0 \in I$ and $t \in [0, T]$ so that $\xi(x_0, t) \in U$?



How to find unsafe trajectories in non-linear systems?

- *Non-linearity* may arise from the system dynamics (i.e. f) or the controller (e.g. when g is a Neural Network).
- The solution to non-linear ODEs do not typically have a closed form, and hence novel tools are needed to analyze safety of such systems.
- We assume access to a forward simulator that estimates the path $\xi(x_0, \cdot) : [0, T] \rightarrow \mathbb{D}$ for each $x_0 \in I$.

Existing tools in this domain

Reachable set analysis

- Flow*, X. Chen et al
- CORA, M. Althoff et al
- Sherlock, Dutta et al
- C2E2, P. S. Duggirala et al
- DryVR, C. Fan et al

Search for unsafe trajectories

- S-TaLiRo, Y. Annpureddy et al
- Breach, A. Donze et al

Outline

1. Guided State-Space Exploration in Closed Loop Systems
    - Introduction to Cyber-Physical Systems (CPS)
    - Verifying safety of CPS
    - State space exploration using sensitivity
    - NExG: state space exploration by learning sensitivity direction
2. Groupwise cross-correlation mining in bi-view data
    - Bi-view data and Bimodules
    - Bimodules as (almost) communities in the cross-correlation network
    - Stable Bimodules and the Bimodule Search Procedure
    - eQTL analysis using GTEx data

How to find unsafe trajectories in non-linear systems?..continued

State space exploration using sensitivity function

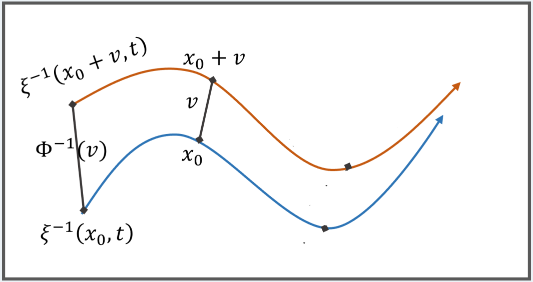
Sensitivity functions

$$\Phi(x_0, v, t) \doteq \xi(x_0 + v, t) - \xi(x_0, t)$$
$$\Phi^{-1}(x_0, v, t) \doteq \xi^{-1}(x_0 + v, t) - \xi^{-1}(x_0, t)$$

where $\xi^{-1}(x, t) = \xi(x, -t)$.

Explore new trajectories using inverse-sensitivity

If we can evaluate $\Phi^{-1}(x_0, v, t)$, then we can find a new curve that passes through the point $x_0 + v$ at time $t > 0$.



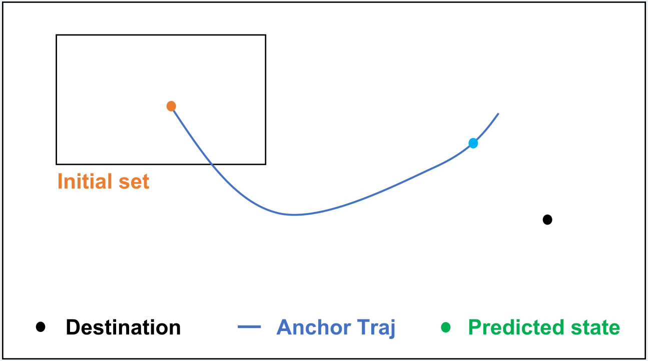
State space exploration using sensitivity function

Reachability Problem

Given destination $z \in \mathbb{D}$ and $t \in [0, T]$, find a $x_0 \in I$ such that $\xi(x_0, t) = z$.

Goyal and P. S. Duggirala, 2020

Use neural networks to learn Φ^{-1} from system traces in order to address reachability.



Using neural networks to learn the sensitivity function

Choice of Neural Network Architecture

- Neural networks are universal approximators [Gühring et al., 2020] that can learn smooth functions based on collections of input-output pairs.
- Although there is increasing theoretical understanding of Deep Networks (see e.g. [Bartlett et al. 2021], [He and Tao, 2021], and [Berner et al. 2021]), training neural networks still remains more of an art than science.
- We used a NN with 3 layers with 512 neurons each (activation RBF for Layer 1 and ReLU for Layers 2 and 3), and an output layer with linear activation.

Generating training sets

From two neighboring trajectories $\{x_{ih}\}_{i=0}^{T/h}$ and $\{x'_{ih}\}_{i=0}^{T/h}$ generate input-output pairs

$$\Phi^{-1}(x_t, x'_t - x_t, t) = x'_0 - x_0 \quad t \in \{h, 2h, \ldots,\}$$

for learning Φ^{-1}.

1. Simulate trajectories starting from $M = 40$ initial (random) points for time T.
2. For each initial point, $L = 10$ trajectories are generated at random starting from starting from a small neighborhood $\|v\| = 0.01$ of the point.

Outline

1. Guided State-Space Exploration in Closed Loop Systems
   - Introduction to Cyber-Physical Systems (CPS)
   - Verifying safety of CPS
   - State space exploration using sensitivity
   - NExG: state space exploration by learning sensitivity direction
2. Groupwise cross-correlation mining in bi-view data
   - Bi-view data and Bimodules
   - Bimodules as (almost) communities in the cross-correlation network
   - Stable Bimodules and the Bimodule Search Procedure
   - eQTL analysis using GTEx data

NExG: Learn only the direction of sensitivity vector for small perturbations

Goyal, Dewaskar and P. S. Duggirala, 2022

- The iterations in Goyal and P. S. Duggirala, 2020 are not guaranteed to converge once they reach a certain neigborhood.
- Reason: the error in NN approximation to $\Phi^{-1}(x_0, v, t)$ does not converge to zero as the perturbation v converges to zero.
- We address this issue in NExG by using neural networks to learn only the direction vector $\tilde{\Phi}^{-1} = \Phi^{-1}/\|\Phi^{-1}\|$ for small perturbations.

- When $\|v\| \ll 1$, by Taylor's expansion

$$\tilde{\Phi}^{-1}(x_0, v, t) \approx \frac{\nabla_v \Phi^{-1}(x_0, 0, t)e}{\|\nabla_v \Phi^{-1}(x_0, 0, t)e\|}$$

where $e = v/\|v\|$.

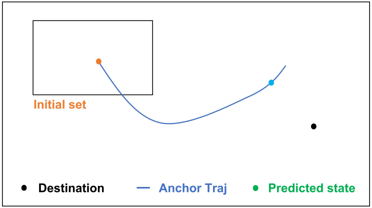


Figure: Reachability using the exact "compass" $\tilde{\Phi}^{-1}$

NExG: Convergence theorem via a contraction argument

Goyal, Dewaskar, and P. S. Duggirala, 2022

Sensitivity approximation error using $N_{\Phi^{-1}}$

$$\|N_{\Phi^{-1}}(x_t, v, t) - \Phi^{-1}(x_t, v, t)\| \leq \varepsilon_{\text{rel}}\|\Phi^{-1}(x_t, v, t)\| + \varepsilon_{\text{abs}}$$

for any $x_t \in \mathbb{D}$, $t \in [0, T]$, and $\|v\| \leq r$.

Assumptions

- $N_{\Phi^{-1}}$ satisfies the above approximation bounds.
- The system satisfies suitable growth bounds, and initial set is unconstrained ($I = \mathbb{D}$.)
- Error coefficients $\varepsilon_{\text{rel}}, \varepsilon_{\text{abs}} > 0$ are sufficiently small (depending on system growth).

Theorem

The distance dist(k) *between the destination and the output after k iterations of the* *Perturbation Algorithm* *using* $N_{\Phi^{-1}}$ *satisfies:*

$$dist(k) \leq (1 - s\gamma_\epsilon)^k dist(0) + c\varepsilon_{abs} \qquad \forall k \geq 1$$

where $s \in (0, 1]$ *is the scaling factor,* $\gamma_\epsilon \in (0, 1)$ *and* c *are values independent of* k*.*

Performance on benchmarks with neural feedback controllers

NExG has better convergence performance compared to Neural Explorer

- We performed comparative analysis with Neural Explorer and S-TaLiRo across 20 benchmark systems with Neural Feedback Controllers (taken from ARCH test suite).
- Relative error for NExG was 1-3% as compared to 10-15% for Neural Explorer, with fewer simulation.

NN Controller

7 layers

100 neurons/layer

$$\dot{x} = y$$
$$\dot{y} = u * y * y - x + w$$

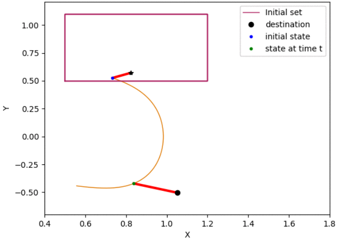


Figure: NExG

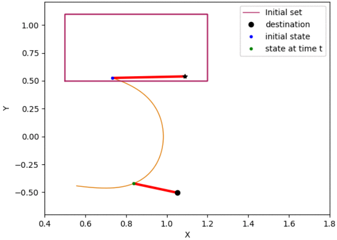


Figure: Neural Explorer [Goyal and P. S. Duggirala, 2020]

Performance on benchmark with neural feedback controllers

NExG features: course-correction period p

One can reduce operational cost by simulating new trajectory less often, a step we call as course-correction.

NN Controller

7 layers

100 neurons/layer

$$\dot{x} = y$$

$$\dot{y} = u * y * y - x + w$$

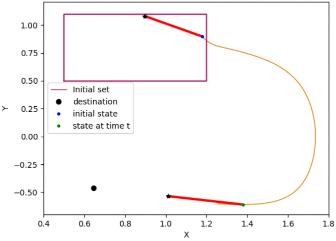


Figure: Trajectory simulation at every iteration ($p = 1$)

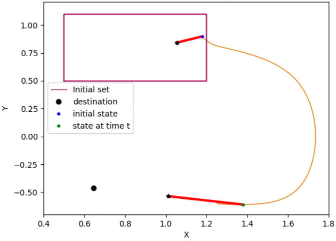
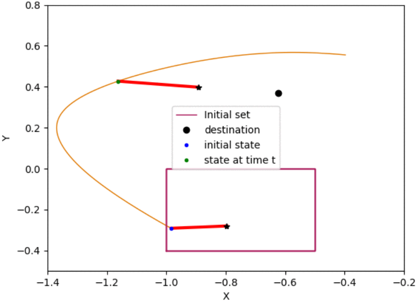


Figure: Trajectory simulation once every 3 iterations ($p = 3$)

Performance on benchmarks with neural feedback controllers

NExG feature: satisfying initial constraint by projection



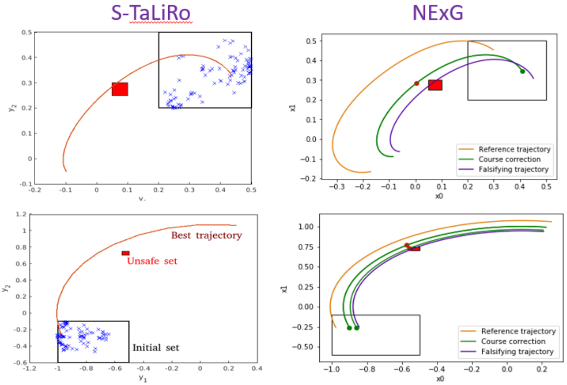
$$\dot{x_1} = x_2$$
$$\dot{x_2} = -x_1 + 0.1 * \sin(x_3)$$
$$\dot{x_3} = x_4$$
$$\dot{x_4} = u \quad \text{←----}$$

NN Controller

3 layers

100 neurons/layer

Application to finding an unsafe counterexample



- achieves considerable performance gain

- can supplement existing falsification tools

Summary: guided state-space exploration in closed-loop control systems using sensitivity approximation

We saw:

- How to explore new system trajectories by learning the sensitivity function for the system.
- NExG learns the sensitivity direction for small perturbations using a neural network based on simulated system traces.
- We theoretically & empirically (on benchmark systems with Neural Feedback Controller) demonstrate state space exploration to assess reachability and to find unsafe trajectories.

Arxiv Manuscript: https://arxiv.org/pdf/2207.03884.pdf

Future directions to address limitations of current work

- Move from benchmarks (currently upto 6D) to Industrial systems (e.g. Auto ACAS F-16)
- Extend convergence guarantees to the case of a constrained initial set, and for tasks beyond reachability – for instance, safety and general MTL specification.
- Rigorously assess the approximation error between true sensitivity function and its Neural Network based approximation.

Outline

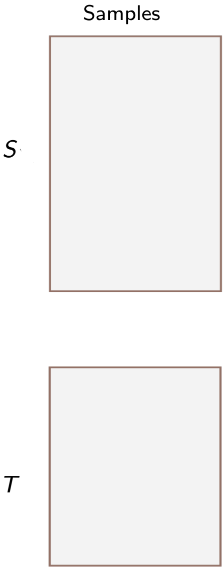
1. Guided State-Space Exploration in Closed Loop Systems
    - Introduction to Cyber-Physical Systems (CPS)
    - Verifying safety of CPS
    - State space exploration using sensitivity
    - NExG: state space exploration by learning sensitivity direction

2. Groupwise cross-correlation mining in bi-view data
    - Bi-view data and Bimodules
    - Bimodules as (almost) communities in the cross-correlation network
    - Stable Bimodules and the Bimodule Search Procedure
    - eQTL analysis using GTEx data

Outline

1. Guided State-Space Exploration in Closed Loop Systems
   - Introduction to Cyber-Physical Systems (CPS)
   - Verifying safety of CPS
   - State space exploration using sensitivity
   - NExG: state space exploration by learning sensitivity direction
2. Groupwise cross-correlation mining in bi-view data
   - Bi-view data and Bimodules
   - Bimodules as (almost) communities in the cross-correlation network
   - Stable Bimodules and the Bimodule Search Procedure
   - eQTL analysis using GTEx data

Bi-view data and a related exploratory problem



Measurements of two types of features

$S = \{s_1, \ldots, s_p\}$ & $T = \{t_1, \ldots, t_q\}$

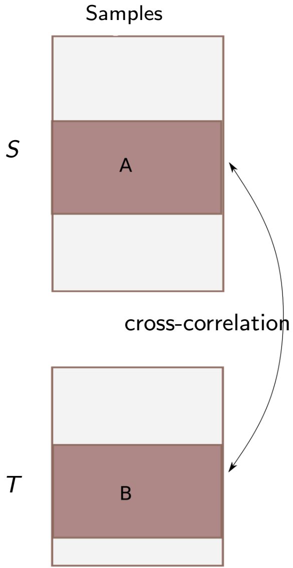
on n common samples. Typically $p, q \geq n$.

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)

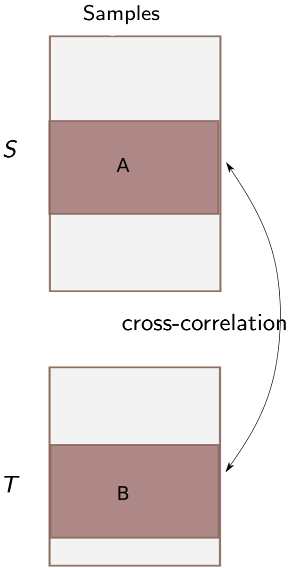
(A, B) is a bimodule if

- $A \subseteq S$ 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

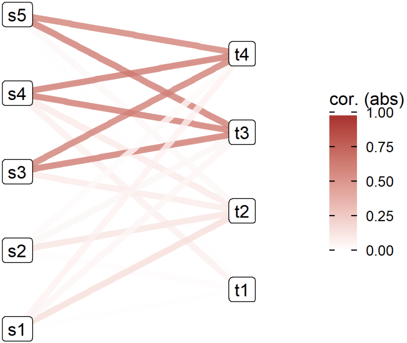
Outline

1. Guided State-Space Exploration in Closed Loop Systems
    - Introduction to Cyber-Physical Systems (CPS)
    - Verifying safety of CPS
    - State space exploration using sensitivity
    - NExG: state space exploration by learning sensitivity direction

2. Groupwise cross-correlation mining in bi-view data
    - Bi-view data and Bimodules
    - Bimodules as (almost) communities in the cross-correlation network
    - Stable Bimodules and the Bimodule Search Procedure
    - eQTL analysis using GTEx data

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)

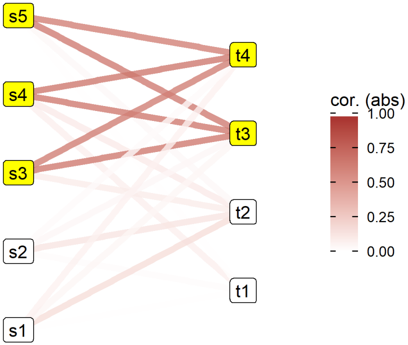
Nodes in a community are more correlated, on average, to nodes inside the community than to nodes outside.

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

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 $B = \{t_3, t_4\}$.

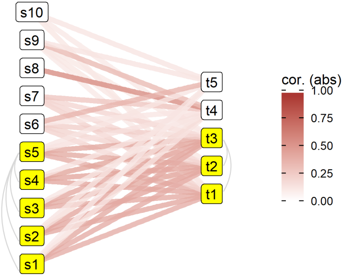
Community (rough definition)

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.)

Almost cross-correlation communities: role of intra-correlations



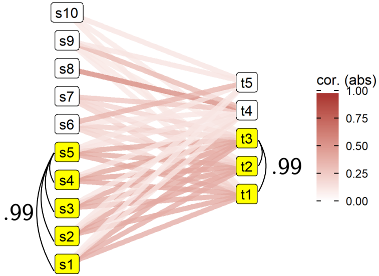
$A = \{s_1, \ldots, s_5\}$, $B = \{t_1, t_2, t_3\}$

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

Likely to see this community by chance in random data?

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

Almost cross-correlation communities: 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.

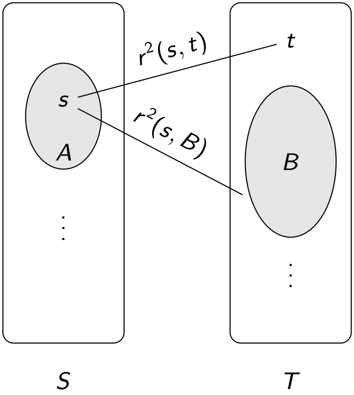
Likely to see this community by chance in random data? Yes

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

Outline

1. Guided State-Space Exploration in Closed Loop Systems
   - Introduction to Cyber-Physical Systems (CPS)
   - Verifying safety of CPS
   - State space exploration using sensitivity
   - NExG: state space exploration by learning sensitivity direction
2. Groupwise cross-correlation mining in bi-view data
   - Bi-view data and Bimodules
   - Bimodules as (almost) communities in the cross-correlation network
   - Stable Bimodules and the Bimodule Search Procedure
   - eQTL analysis using GTEx data

Stable Bimodules



Stable bimodule (definition)

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

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

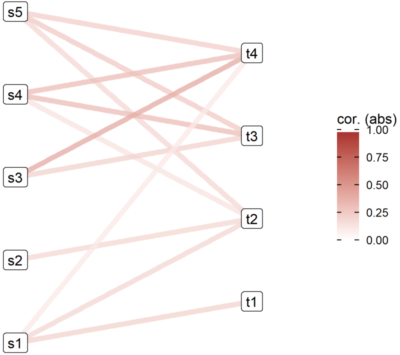
- Recursive definition like a community; made precise using hypothesis testing (details).
- Permutation test accounts for intra-correlations.
- Benjamini-Yekutieli correction for multiple testing.
- Interested in connected stable bimodules

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: example run

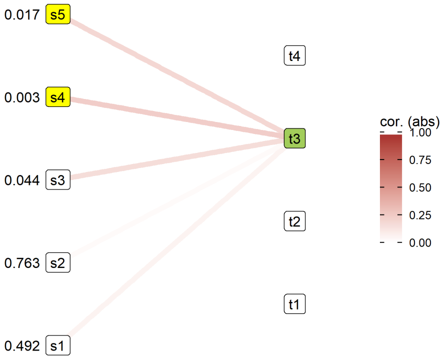


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

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

Stable bimodule found.

Bimodule Search Procedure: example run

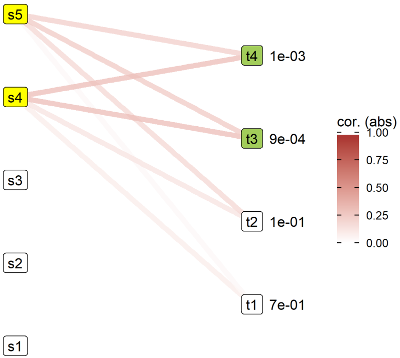


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

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

Stable bimodule found.

Bimodule Search Procedure: example run

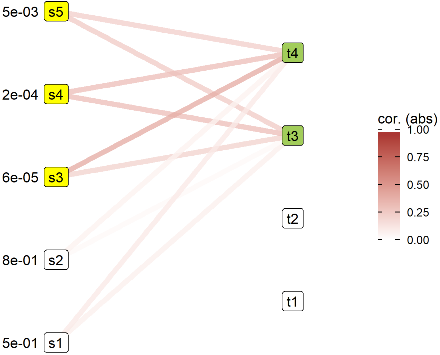


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

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

Stable bimodule found.

Bimodule Search Procedure: example run

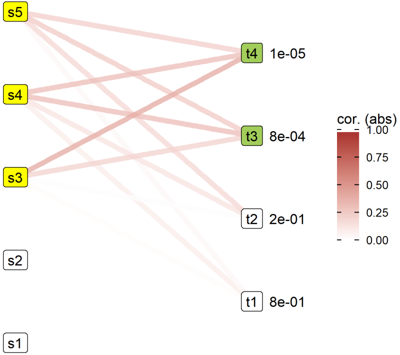


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

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

Stable bimodule found.

Bimodule Search Procedure: example run

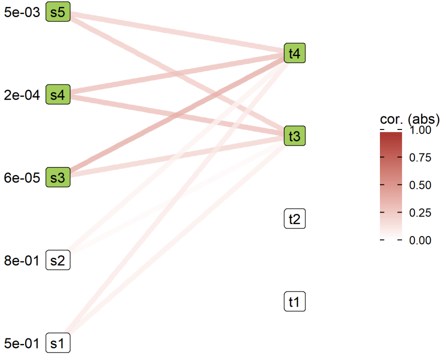


1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

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

Stable bimodule found.

Bimodule Search Procedure: example run



1. $B_0 = \{T_3\}$
2. $A_0 = \{S_4, S_5\}$
3. $B_1 = \{T_3, T_4\}$
4. $A_1 = \{S_3, S_4, S_5\}$
5. $B_2 = \{T_3, T_4\}$
6. $A_2 = \{S_3, S_4, S_5\}$

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

Stable bimodule found.

Outline

1. Guided State-Space Exploration in Closed Loop Systems
   - Introduction to Cyber-Physical Systems (CPS)
   - Verifying safety of CPS
   - State space exploration using sensitivity
   - NExG: state space exploration by learning sensitivity direction

2. Groupwise cross-correlation mining in bi-view data
   - Bi-view data and Bimodules
   - Bimodules as (almost) communities in the cross-correlation network
   - Stable Bimodules and the Bimodule Search Procedure
   - eQTL analysis using GTEx data

Data from GTEx project (v8)

from gtexportal.org

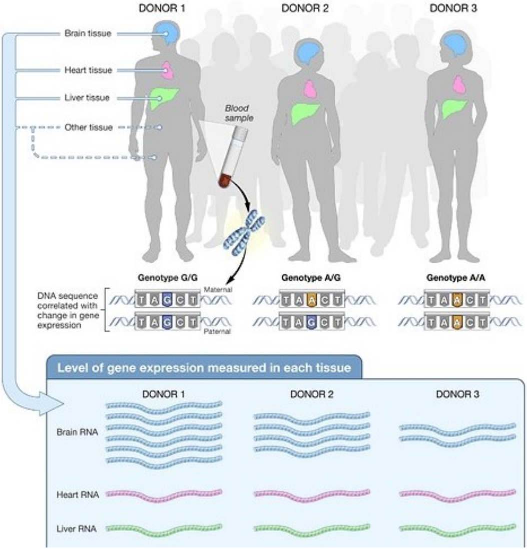
NIH funded GTEx project
A large collection of multi-tissue eQTL data from donors.

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

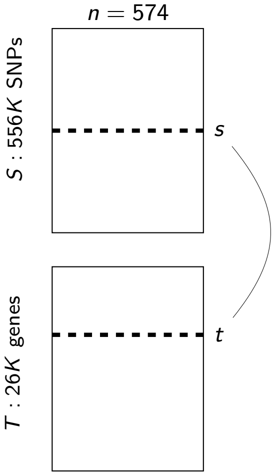
Expression measured in multiple tissues
RNA sequencing used to measure expression of genes.

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 (limits statistical power).

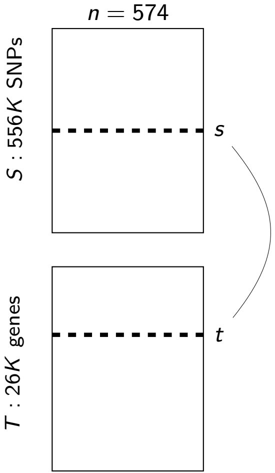
Groupwise eQTL: find SNP-gene bimodules (CONDOR)

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 Hence bimodules may represent a group of SNPs that disrupt the functioning of gene regulatory networks and contribute to diseases

We use the Bimodule Search Procedure!

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 (limits statistical power).

Groupwise eQTL: find 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 Hence bimodules may represent a group of SNPs that disrupt the functioning of gene regulatory networks and contribute to diseases

We use the Bimodule Search Procedure!

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.

Summary: groupwise cross-correlation mining in bi-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 Bimodule Search Procedure. Extensions to multi-view data and other types of correlations.

Appendix

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.

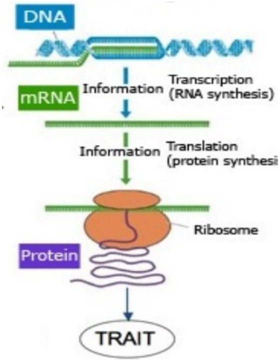
- Almost all (99.3%) bimodules have at least one local SNP-gene pair.
- In addition, almost half of the larger bimodules found gene and SNPs that had distal effects.

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?

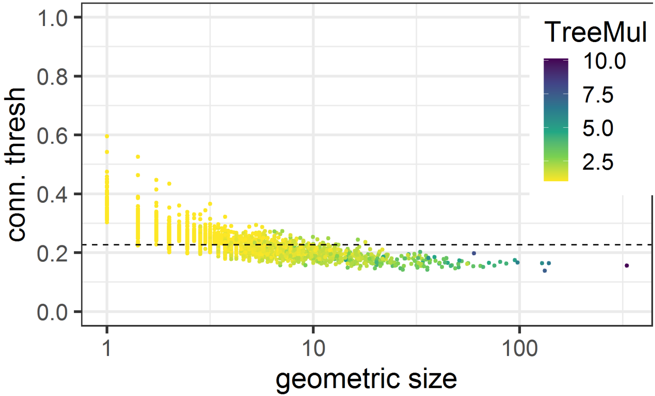
Conservative estimate of strongest edges

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.

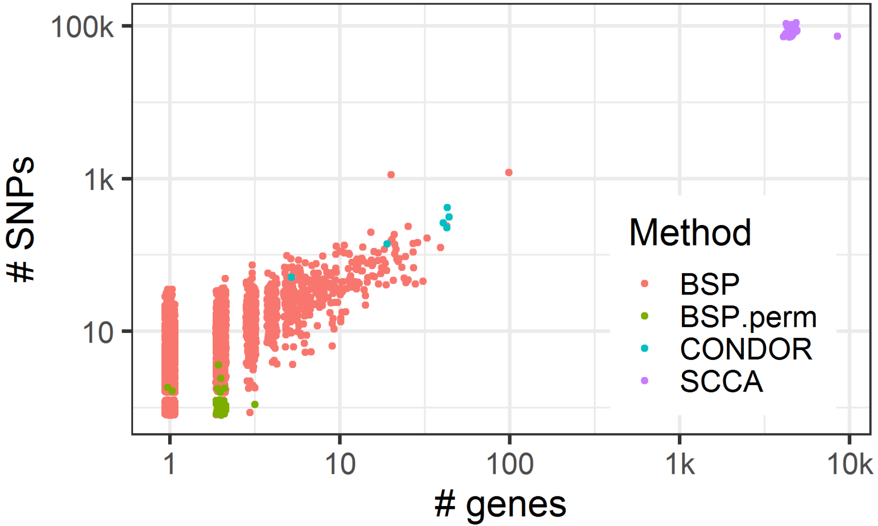
Enrichment of known gene sets in bimodules

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?

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

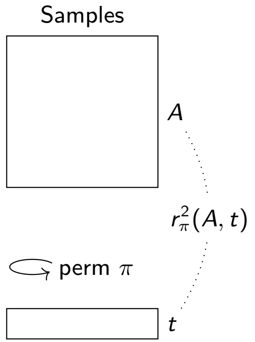
Steps

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

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

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

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



Samples

A

$r^2_\pi(A,t)$

perm π

t

Permutation p-value

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

Fast computation + other details

Hypothesis testing implementation details

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

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

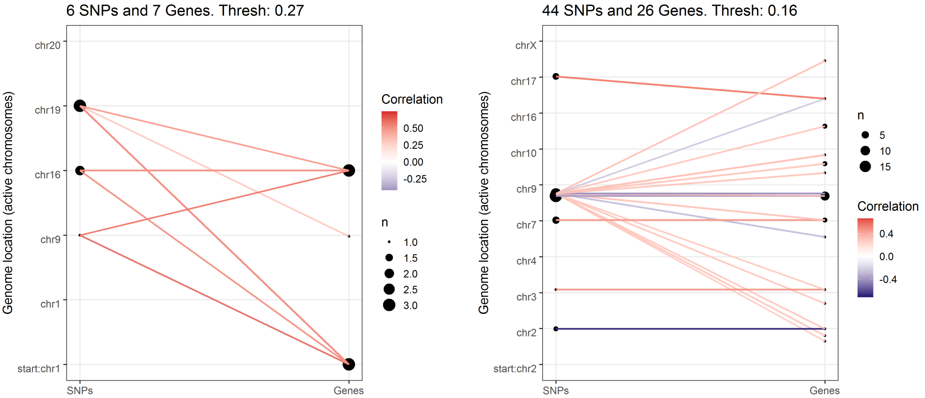
which conditions on correlations in A.

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.

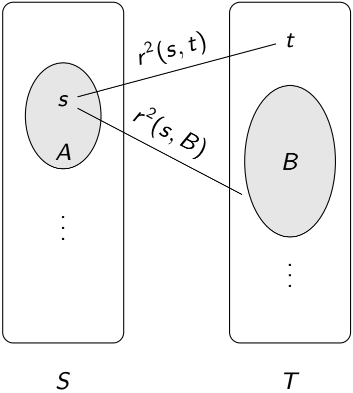
New potential eQTLs from bimodules

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.

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 $A_0 \subseteq S$.

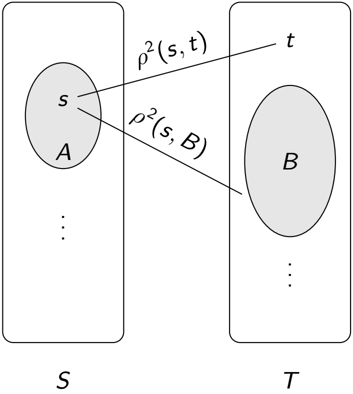
Bimodule Search Procedure (BSP)

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

References

Patricia Derler, Edward A Lee, and Alberto Sangiovanni Vincentelli. Modeling cyber–physical systems. *Proceedings of the IEEE*, 100(1):13–28, 2011.