CSB Primer

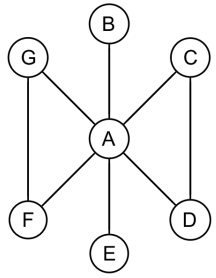
Let's learn some CBB Crew

January 2020

Topic	Assigned
Graph Theory	Gian
Probabilistic Graphical Models	Gian
Stoichiometric Network Analysis	Martin
Dynamic Systems	
Systems Identification	
Simplified Dynamic Models	Samuel
Stochastic Systems	Gian

1 Graph Theory

1.1 Adjacency Matrix



For a graph of n nodes the adjacency matrix has size $n \times n$, non-zero entries of the matrix represent the connections between two neighboring nodes.

Example
The adjacency matrix of the sample graph in upper triangular form is given by:.

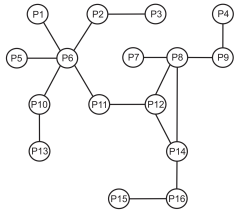
$$\begin{pmatrix} - & A & B & C & D & E & F & G \\ A & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ B & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ C & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ E & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ F & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ G & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

1.2 Degree Distribution

The degree distribution represents the number of nodes with a given degree (number of direct neighbours). The average degree is a metric for the connectedness of the whole graph.

$$\overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

The degree distribution is also used to find the graph family (e.g. scale-free, random, small world) of a given network.



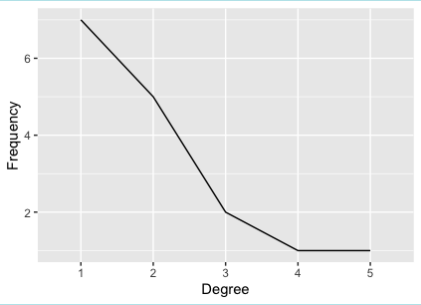
Example
The following table gives the degrees for each node in the graph above.

P1	1	**P9**	2
P2	2	**P10**	2
P3	1	**P11**	2
P4	1	**P12**	3
P5	1	**P13**	1
P6	5	**P14**	3
P7	1	**P15**	1
P8	4	**P16**	2

Counting the frequency of each degree gives the following table.

Degree	Frequency
1	7
2	5
3	2
4	1
5	1

Plotting Degree-Frequency gives the following plot.



This indicates that the network is scale-free; if the network was random the distibution would be Poisson.

1.3 Clustering Coefficient

The clustering coefficient is a measure for how connected with each other neighbours of a particular node are.

$$C = \frac{2e}{k_u(k_u - 1)}$$

Where e is the number of edges inbetween neighbours and k_u is the number of neighbours.

Example
For the network given in subsection (1.1) we get the following:

$$C = \frac{2e}{k_u(k_u - 1)} = \frac{2 \times 2}{6 \times (6 - 1)} = \frac{2}{15}$$

The network global clustering coefficient if given by the following equation.

$$\bar{C} = \frac{1}{n} \sum_{i=1}^{n} C_i$$

1.4 Cliques

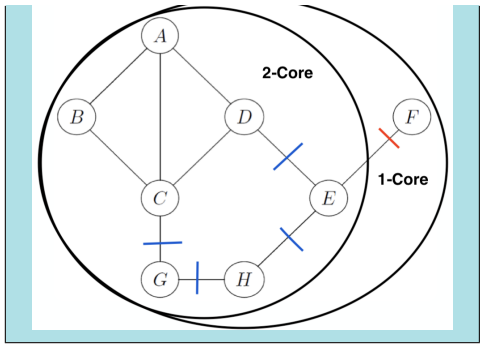
A **clique** is a subset of nodes such that every two nodes of that subset are connected. Thus this subsets of nodes represents a complete graph.
A **maximal clique** is one that cannot be extended, as no other neighbours of any node are connected to every node of the clique.
A **maximum clique** is the largest clique within a given network.

1.5 K-Cores

The k-core if a graph G is the maximal subgraph $H \subseteq G$ such that $\delta(H) \geq k$, indicating that every node of H has at least k neighbours. Thus the 0-core includes all nodes and the 1-core includes all but isolated nodes.
General Procedure

```
For every k until the graph is an empty set
    Until no more nodes are removed
        Remove all nodes with deg(N) <= k;
        these nodes belong to the k-core
```

Example
The k-cores for the following graph are determined by removing all edges with degree $deg(N) = 1$ in a first step. This is only node F, thus F is removed and labeled with core 1. Now all nodes with $deg(N) \leq 2$ are removed: $\{G, H, E\}$. The remaining nodes now also only have two neighbours each, all edges are removed and the remaining nodes are part of the 2-core.



2 Probabilistic Graphical Models

2.1 Metropolis Hastings

2.1.1 A Markovian Tale

"We are employed as a contractor for a mining company to map the amount of subterranean iron across a vast, lifeless desert. The desert is flat and uninformative of the treasures that lie underneath. However, fortunately, we have a machine that measures the magnetic field directly underneath, which varies in direct proportion to the total amount of iron below. Suppose that the mining company has already determined that the area is rich with iron deposits and is interested only in mapping the relative abundance of deposits over the desert. How should we approach mapping the underground iron? The simplest way would be to survey the magnetic field at, say, 1km intervals. However, even at this modest resolution, we would need to sample 1000 1000 = 1million points. If instead we increased the precision to 100 metres, we would then need to take 100 million samples. We'd die of thirst! There must be a quicker way to build an accurate map. Suppose that we start in a random location in the desert and measure the magnetic field beneath. We then use a random sample from a bivariate normal distribution centred on our current location, to pick a new location to sample. We then measure the magnetic field there, and if it exceeds the value at the old site, we move to the new location and add the new (north, east) location to our list. By contrast, if the value of the magnetic field is lower than the current value, then we only move there probabilistically, with a probability given by the ratio of the new value to the old. To do this we compare the ratio with a random sample from a uniform distribution, p U(0,1). If our ratio exceeds p, then we move there and add the new (north, east) to our current list. If it does not, then we move back to where we were, and add our previous location to our list again."

Excerpt From: Ben Lambert. "A Student's Guide to Bayesian Statistics"

2.1.2 Motivating Metropolis Hastings

Suppose we want to sample from a distribution which follows a known function p(x), but for this function we cannot obtain the corresponding probability density function. The probability density function has the following property.

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

For simple function the normalizing constant is easy to obtain. Suppose we have the following function.

$$p(x) = e^{-x^2/2}, x \in (-\infty, \infty)$$

This yields:

$$\int_{-\infty}^{\infty} p(x)\,dx = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}$$

So we normalize our function in the following manner.

$$\int_{-\infty}^{\infty} \varphi(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,dx = 1$$

So for this simple example we found the normalizing factor to be $\sqrt{2\pi}$. But for complicated functions it can be very difficult to find the normalizing constant and therefore the corresponding probability density function.

2.2 Conditional Probabilities

Conditional probability is a measure of the probability of an event occurring given that another event has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

These probabilities are found by considering the probability tables.

2.3 Conditional Independence

Two random events A and B are conditionally independent given a third event C precisely if the occurrence of A and the occurrence of B are independent events in their given C.
X is conditionally independent from Y given Z ($X \perp Y|Z$) if the following holds.

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

where we can use the following equality according to the conditional probability:

$$P(X, Y|Z) = \frac{P(X \cap Y \cap Z)}{P(Z)}$$

This also works for unconditioned probabilities, if the following holds X and Y are independent.

$$P(X \cap Y) = P(X)P(Y)$$

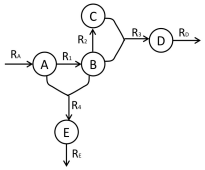
2.4 Law of Total Probability

$$P(A) = \sum_n P(A \cap B_n \cap C_n \cap ...)$$

3 Stoichiometric Network-Analysis

3.1 Stoichiometric Matrix

The stoichiometric matrix (N) has dimension n (number of internal metabolites) x q (number of metabolic reactions)



Example
The following matrix gives the stoichiometric matrix to the network above. The corresponding matrix is:

$$\begin{pmatrix} - & Ra & Rd & Re & R1 & R2 & R3 & R4 \\ A & 1 & 0 & 0 & -1 & 0 & 0 & -1 \\ B & 0 & 0 & 0 & 1 & -1 & -1 & -1 \\ C & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ D & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ E & 0 & 0 & -1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

3.2 Flux Distribution

The flux distribution r describes all feasible fluxes, whereas there is a feasibility criterion ($r_i \geq 0$).

3.3 Mass Balance

All internal metabolites are in balance, meaning that the fluxes for each species are the sum of all going in minus the ones going out.

$$\frac{dc_i}{dt} = fluxes_{i,in} - fluxes_{i,out} \quad (1)$$

3.4 Balancing Equation

The balancing equation is the product of the stoichiometric matrix and the flux distribution vectors.

$$\frac{dc_i}{dt} = N * r(t) \quad (2)$$

N is invariant whereas r is the time variant. Under (quasi) steady state conditions we assume the rate of change to be zero :

$$N * r = 0 \quad (3)$$

3.5 Kernel

All feasible solutions lie in the nullspace of N with dimensions

$$d = q - rank(N) \quad (4)$$

The rank describes the linearly independent column vectors of N. (get the Matrix into row echelon form and look at the non-zero rows of the transpose). Any solution r is given by a linear combination of the columns of k; k is the basis of the solution space.

3.6 Conservation Relations

Weighted sums of metabolite concentrations that are always the same. Correspond to the linearly dependent rows in N. They lie in the left null space of N - the derivative is zero. Therefore we can find CRs by using the following formula:

$$y^T * N = 0 \; or \; y * N^T = 0 \tag{5}$$

Example
The following example shows the way to show what are CRs:

$$2[A] - [B] + [C] = 0 \tag{6}$$
$$-[A] + [B] + [D] \tag{7}$$
$$N * y^T = 0 \tag{8}$$
$$\forall N = \begin{pmatrix} -1 & -1 & 1 & 0 \\ 0 & -1 & -1 & 1 \end{pmatrix} \tag{9}$$
$$y^T = \begin{pmatrix} 2 \\ -1 \\ 2 \\ 0 \end{pmatrix} \vee \begin{pmatrix} -1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \tag{10}$$
$$\tag{11}$$

3.7 Flux Balance Analysis

Incorporate further constraints to limit the network behaviour (Quasi steady state, reaction reversibilities, optimal feasibility criterion).

3.8 Flux Variability Analysis

Identify minimal and maximal fluxes. Then identify the maximum of one objective as the new constraint. Return the min and max of what is possible. Has the problem of running into a pareto optimum.

4 Dynamic Systems

4.1 Stability Analysis of Steady States

1. Calculate the Jacobian from the differential equations
2. Plug in the steady state values for x and y
3. Check if $tr(J_{SS}) < 0$ and $det(J_{SS}) > 0$ as conditions for stability and $tr^2 - 4det < 0$ as a condition for oscillations

5 Systems Identification

5.1 Sensitivity Matrix

Indicates how strongly the system responds in changes to different parameters.
ADD SENSITIVITY MATRIX

5.2 Fisher Information Matrix

$$F(p) = \sum_{i=1}^{N} \left(\frac{S(t_i)^T S(t_i)}{\sigma(t_i)} \right), \sigma_i \geq \sqrt{[F(p)^{-1}]_{ii}}$$

A 2×2-matrix is inverted as follows.

$$\mathbf{A}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

5.3 Maximum Likelihood Parameter Estimates

We define the objective function; minimizing the error between the model and the experimental data.

$$\phi(P) = \sum_{i=1}^{n} \left(\frac{e(t_i)}{\sigma(t_i)} \right)^2$$

Where $e(t_i) = x(t_i) - x_m(t_i)$
Find the minimal value by calculating the partial derivatives $\frac{\partial \phi}{\partial x_0}$ and $\frac{\partial \phi}{\partial k}$ of the function. Solve the resulting equations for x_0 and k.

6 Simplified Dynamic Models

6.1 Qualitative dynamics

Nullclines First we want to find the nullclines. Let's suppose we have two ODEs for X_a and X_b solve them for 0 and express these as functions of X_a and X_b.

$$\frac{dX_a}{dt} = 0 \quad \frac{dX_b}{dt} = 0 \tag{12}$$

Steady states Now we want to find one ore more steady state(s). Again we solve the ODEs for 0 but now we set them equal to get rid of the parameters X_a, X_b. We do this by expressing one function in terms of X_a and plug this into the other function. The coordinates of the steady states should be expressed as a function of the unknown parameters e.g. the reaction rates k_a and k_b.

Qualitative phase portrait Using the nullclines, the steady state and a vector field we will see the behaviour of our system in the qualitative phase portrait. To get the **vector field** we have to look under which conditions our ODE systems (dX_i/dt) is greater or equal to 0. This corresponds to in which direction our system changes. Let's suppose that X_a is on the x-axis and X_b is on the y-axis in our phase portrait. If $dX_a/dt > 0$ we move towards higher values of X_a and if $dX_a/dt < 0$ we move towards smaller values of X_a in the direction of the x-axis. We do the same for X_b and combine the information to a vector of movement. The nullcline corresponds to zero change, so if we cross a nullcline the corresponding direction of the vector should change. The intersection(s) of the nullclines corresponds to the steady state(s).

Derivative sign pattern The derivative sign pattern

$$\pi(R_{\text{Region}}) \quad \text{Region} = 1, ..., n \tag{13}$$

shows in which direction the vector field "moves" in different regions. To get the derivative sign pattern we split the graph into different regions. Every area that lies between nullclines is a region, the nullclines are separate regions and if there is an intersection with another nullcline every sub part of each nullcline is a unique region and the intersections are a region. The derivative sign pattern is a vector with dimensions of the phase portrait (number of ODEs). The entries of the vector have a $+$ sign, if the corresponding $dX_i/dt > 0$, a $-$ sign, if the corresponding $dX_i/dt < 0$ or a 0, if the corresponding $dX_i/dt = 0$.

Example

6.2 Logical Models

Boolean 0 means Gene inactive $\rightarrow$ Protein absent, boolean 1 means Gene active $\rightarrow$ Protein present. The logical functions are:

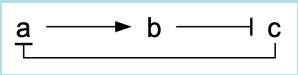
- and: $\wedge$
  $1 \wedge 1 = 1; 1 \wedge 0 = 0; 0 \wedge 0 = 0$
- or: $\vee$
  $1 \vee 1 = 1; 1 \vee 0 = 1; 0 \vee 0 = 0$
- not: $\neg$
  $\neg 1 = 0; \neg 0 = 1$

Logical functions and state table Logicals functions define the output of a module (gene) in the next round given its inputs in the current round. All possible combinations can be visualise in a state table.

Transition graphs If we consider a synchronous update, every input will be evaluated at the same time. We can simply connect every inputs from the state table with its corresponding output from the state table.
When considering asynchronous update we have to evaluate every input separately, take its output as the input for the other rules and check if we end up in a different state.

Steady states Steady state(s) are given by the states that point to themself(s).

Example
Consider the following gene network:

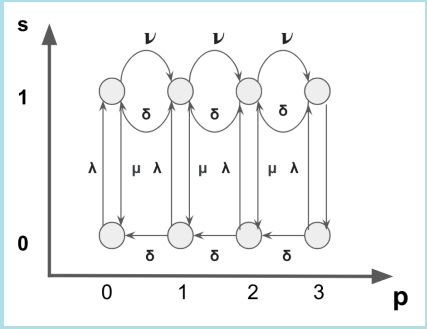


The logical functions are given by:

$$X_a = \neg c, X_b = a, X_c = \neg b$$

7 Stochastic Systems

Example
Write down the chemical master equation of the following system.



We see that this system has two states $s = 1$ and $s = 0$. We formulate chemical master equation for both of these states.

The differential probability of being in state $(s = 0)$ over time is the following.

$$\frac{dP(s=0,n_p,t|s^0,n_p^0,t_0)}{dt} = \underbrace{-(\lambda+\delta n_p)P(s=0,n_p,t|s^0,n_p^0,t_0)}_{①} + \underbrace{\delta(n_p+1)P(s=0,n_p+1,t|s^0,n_p^0,t_0)}_{②} + \underbrace{\mu P(s=1,n_p,t|s^0,n_p^0,t_0)}_{③}$$

① Describes the situation of moving away from any state in $(s = 0)$ to any other state in $(s = 0)$ (Which is expressed by $\delta n_p P(...)$ as this can happen for any of the n_p proteins with rate δ) or to any state $(s = 1)$ (Which is expressed by $\lambda P(...)$, this is independent from n_p as we consider the gene here and not the n_p proteins.) The term is denoted with a leading minus-sign as we describe moving away from said states here.

② Describes the situation of moving from any state $(s = 0, n_p)$ to the state $(s = 0, n_p - 1)$. This can happen for every protein $n_p + 1$. (We had one protein more before this transition), thus we get $\delta(n_p + 1)$ times the probability of being in state $P(s = 0, n_p + 1)$.

③ Describes the situation of moving from any state $(s = 1)$ to any state $(s = 0)$. This transition from the active gene to the passive gene happens with rate μ times the probability of being in state $P(s = 1)$.

Poincaré Diagram: Classification of Phase Portaits in the $(\det A, \operatorname{Tr} A)$-plane