Computing similarity between multiscale biological systems under uncertainty

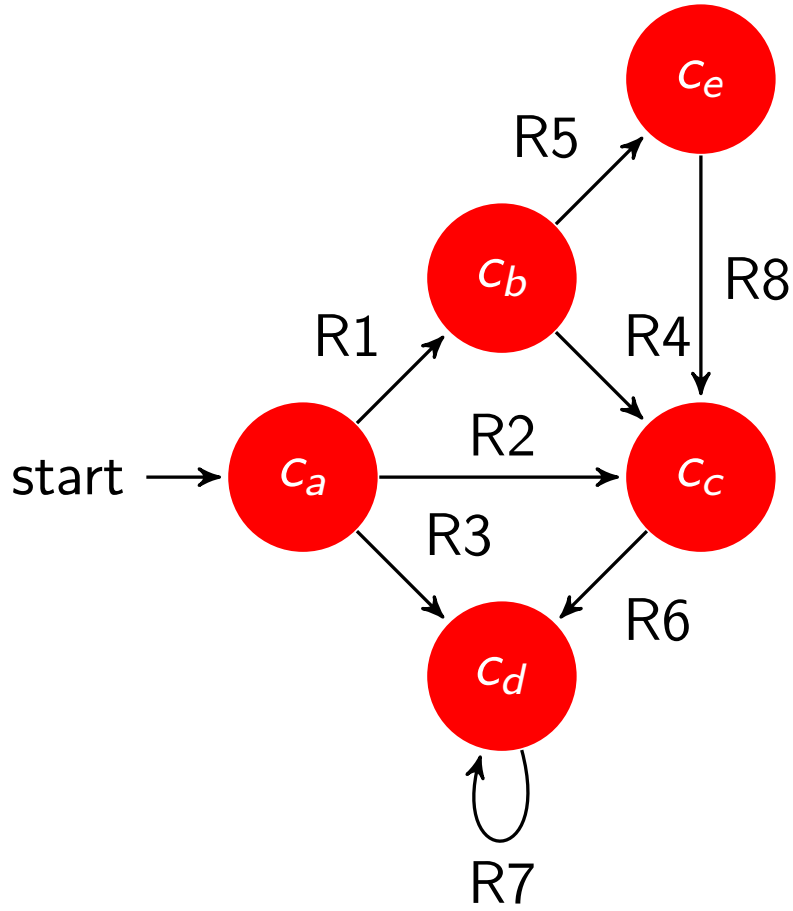
Kris Ghosh

Miami University, Ohio

The Eighth International Workshop on Static Analysis for Systems Biology

An Example: Model of chemical reactions

Reactions: R1,R2,R3,..R8.
Concentration of chemicals:
$c_a, c_b, \ldots, c_e$.



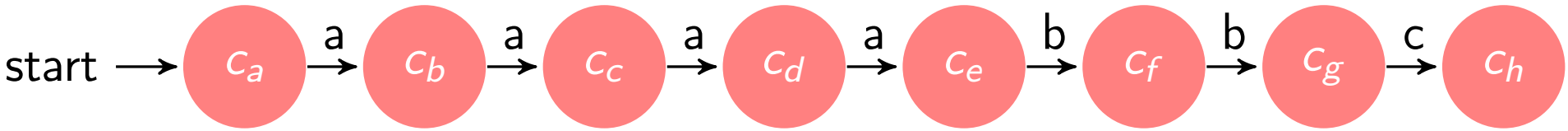
Finite State System
representing chemical reactions

Motivation: Scales in models

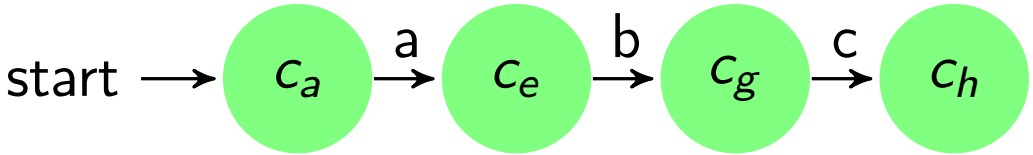
Why multiscale model? Biological systems need an integration of different scales-cellular, molecular , atomic levels...

Motivation: Scales in models

Why multiscale model? Biological systems need an integration of different scales-cellular, molecular , atomic levels...



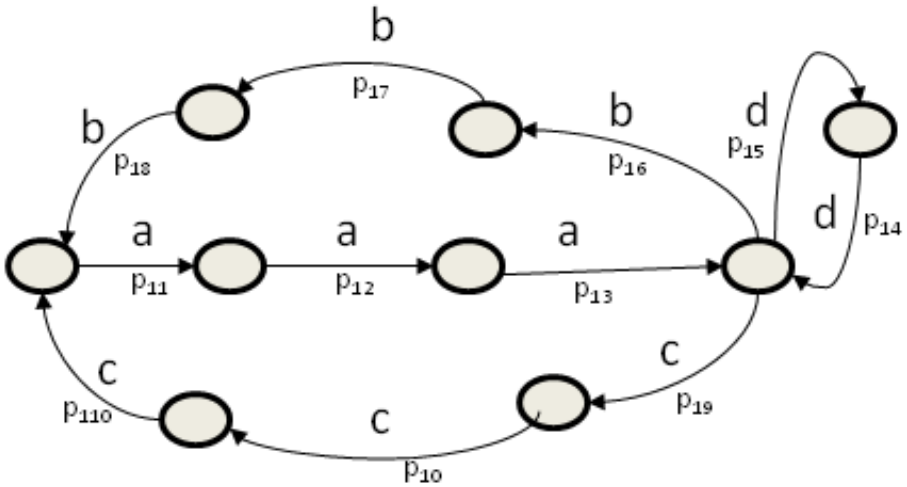
(A)



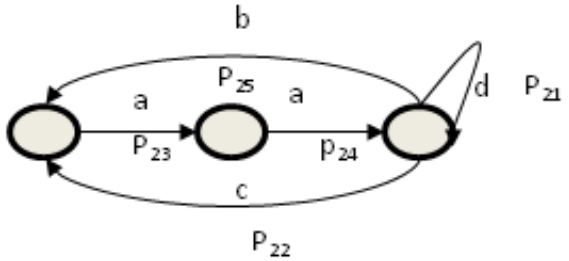
(B)

a,b and c are biological processes.

Multiscales to Discrete Time Markov Chains



(A)



(B)

Discrete Time Markov Chains representing identical partial ordering of pathways represented by edge labels a,b,c and d, respectively. p_{ij} represent the probabilities on the edges where $i, j \in \mathbb{N}$.

Challenges in Modeling in Biology

- Imprecise information
- Incomplete information

Challenges in Modeling in Biology

- Imprecise information
- Incomplete information

A potential solution could be to use: nondeterminism and using probabilistic models.

Challenges in Modeling in Biology

- Imprecise information
- Incomplete information

A potential solution could be to use: nondeterminism and using probabilistic models.
Computational challenge: Large state space of the models.

Outline

Preliminaries

Definition

(Kripke structure) Given a set of propositions, AP, a *Kripke structure*, $\mathcal{K} = \langle S, S_0, E, L \rangle$ consists of

1. S is the set of states.
2. $S_0 \subseteq S$ is the initial set of states.
3. $E \subseteq S \times S$ is the transition relation.
4. $L : S \to 2^{AP}$ where L is the labeling function that labels each state with a subset from the set, AP.

Stuttering Equivalence on Paths

Two infinite paths in Kripke structure $\mathcal{K}$, $\mu = s_o \overset{\alpha_0}{\rightarrowtail} s_1 \overset{\alpha_1}{\rightarrowtail} s_2 \ldots$ and $\nu = r_0 \overset{\beta_0}{\rightarrowtail} r_1 \overset{\beta_1}{\rightarrowtail} \ldots$ are stuttering equivalent ($\equiv_s$)ent if there are two infinite ordered sequences of positive integers, $i = 0 < i_0 < i_1 < \ldots$ and $j = 0 < j_0 < j_1 < \ldots$ such that $\forall k \geq 0$
$L(s_{i_k}) = L(s_{i_k+1}) = \ldots = L(s_{i_{k+1}-1}) = L(r_{j_k}) = L(r_{j_k+1}) = \ldots = L(r_{j_{k+1}-1})$.
The indices i_k and j_k are the starting points of μ and ν, respectively.

Stuttering Equivalence

Two Kripke structures $\mathcal{K}$ and $\mathcal{K}'$ are stuttering equivalent iff

1. The initial states of $\mathcal{K}$ and $\mathcal{K}'$ are the same.
2. For all paths, μ from an initial state, $s_0 \in S_0$ of $\mathcal{K}$, there exists a path ν of $\mathcal{K}'$ from the same initial state of s_0 such that $\mu \equiv_s \nu$.
3. For all paths, ν from an initial state of $s_0 \in S_0$ of $\mathcal{K}'$, there exists a path μ of $\mathcal{K}$ from the same initial state of s such that $\nu \equiv_s \mu$.

Ref: Clarke, E.M., Grumberg and Peled, Model Checking.

Theories

- Interleaving asynchronous
  Ref: Clarke et al, State space reduction using partial order techniques, 1999.
- Bounded Asnchrony
  Ref: J. Fisher et al Bounded asynchrony: Concurrency for modeling cell-cell interactions, 2008.
- Computing bisimulation on structures
  Ref: Paige et al Three partition refinement algorithms, 1987.
- Kullback Leibler Divergence in Systems Biology
  Ref: Petrov Formal reductions of stochastic rule-based models of biochemical systems, 2013.
- Model Reduction in Systems Biology
  Ref: Feret et al, Lumpability Abstractions of Rule-based Systems, 2012.

Labeled transition system (LTS)

Given a set of propositions, AP being the set of labels for states and EL, a set of labels for edges a *labeled state transition system* is defined as $\mathcal{M} = \langle S_0, S, E, L_e, L \rangle$ where,

1. $\langle S, S_0, E, L \rangle$ forms a Kripke structure.
2. $L_e : E \mapsto EL$ is an edge-labeling function.

Labeled Probabilistic System (LPS)

a LPS is a tuple, $\mathcal{W} = \langle S, S_0, \iota_{init}, P, L_e, L, E \rangle$ where:

- $\langle S, S_0, \iota_{init}, P, L \rangle$ is DTMC.
- $L_e : S \times S \to E$ where, E is the set of edge labels.

Measures on Probability Distributions

- Kullback Leibler Divergence (KLD) of two distributions: $H(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$.

- Jensen-Shannon Divergence (JSD) is symmetric verson of KLD: $JSD(P \mid\mid Q) = \frac{1}{2}H(P \mid\mid M) + \frac{1}{2}H(Q \mid\mid M)$ where $M = \frac{1}{2}(P + Q)$.

KLD can only be computed on same state space.

Formalization of System

Read

For an infinite path, $\pi = e_0, e_1, e_2, e_3, \ldots$ in a LPS $\mathcal{W}, (\alpha_0, \alpha_1, \alpha_2, \ldots)$ is the sequence of reaction labels in π. The read of a path is the subsequence of *reaction* labels $\tilde{\pi} = \alpha_0, \alpha_{i_1}, \alpha_{i_2}$ where $0 \leq i_1 \leq i_2 \leq \ldots$, α_{i_j} is in $\tilde{\pi}$ iff $\alpha_{i_j} \neq \alpha_{i_{j-1}}$ and $\alpha_0 \neq \alpha_{i_1}$.

A finite path segment $\sigma = e_0 \rightarrowtail e_1 \rightarrowtail e_2 \rightarrowtail e_3 \ldots \rightarrow e_m \rightarrowtail \ldots$, is identically labeled (*il*) if the reactions are identical. We explicitly allow $m = 0$; in that case we write $e_0 \rightsquigarrow e_0$

Compact Probability on Paths

$P_c(e, e')$ between two edges is computed by the following equations dependent on the label of the successive edge.

$$P_c(e, e') = \begin{cases} P(e, e') \text{ if,} e \neq e' \\ P(e) \times P(e_1) \times \cdots P(e_k) \text{if } e \rightarrowtail e_1 \rightarrowtail \ldots, e_k \rightarrowtail e' \end{cases}$$

The compact probability for an *il* path fragment is computed by the products of the probabilities.

Read equivalence on paths

Paths $\pi_1 \in \Pi(\mathcal{W}_1), \pi_2 \in \Pi(\mathcal{W}_2)$ are read equivalent iff their reads are identical and denoted by $\pi_1 \equiv_r \pi_2$.

Read equivalence on reactions

Given two LPSs, $\mathcal{W}_1$ and $\mathcal{W}_2$, the relation read on edges ($\equiv_r$) is defined on reaction labels, $e_1 \in E_1$ and $e_2 \in E_2$. $e_1 \equiv_r e_2$ if and only if the following conditions hold:

1. $L_(e_1) = L_e(e_2)$.
2. For all paths, $\pi_{e_1} \in \Pi(e_1)$ $\exists$ a path $\pi_{e_2} \in \Pi(e_2)$ such that $\pi_{e_1} \equiv_r \pi_{e_2}$.
3. For all paths, $\pi_{e_2} \in \Pi(e_2)$ $\exists$ a path $\pi_{e_1} \in \Pi(e_1)$ such that $\pi_{e_1} \equiv_r \pi_{e_2}$.

Problem Statement

Given two LTSs, $\mathcal{M}_1$ and $\mathcal{M}_2$, construct two LPSs,$\mathcal{W}_1$ and $\mathcal{W}_2$

- Are $\mathcal{W}_1$ and $\mathcal{W}_2$ read equivalent?
- If yes, compute the KLD between the two structures.

Important:

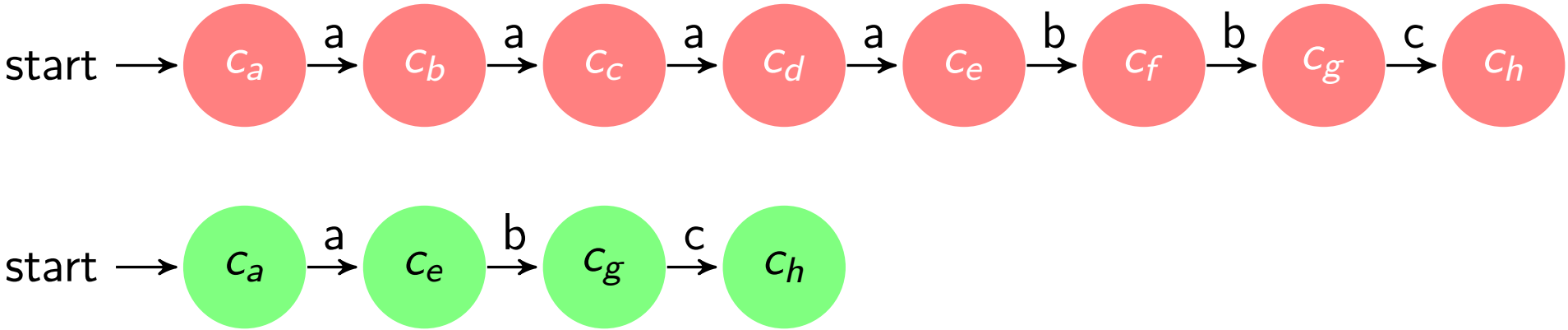
- Only takes account of information of the edges.
- Compares the partial ordering of the two LPSs based on the edge labels.

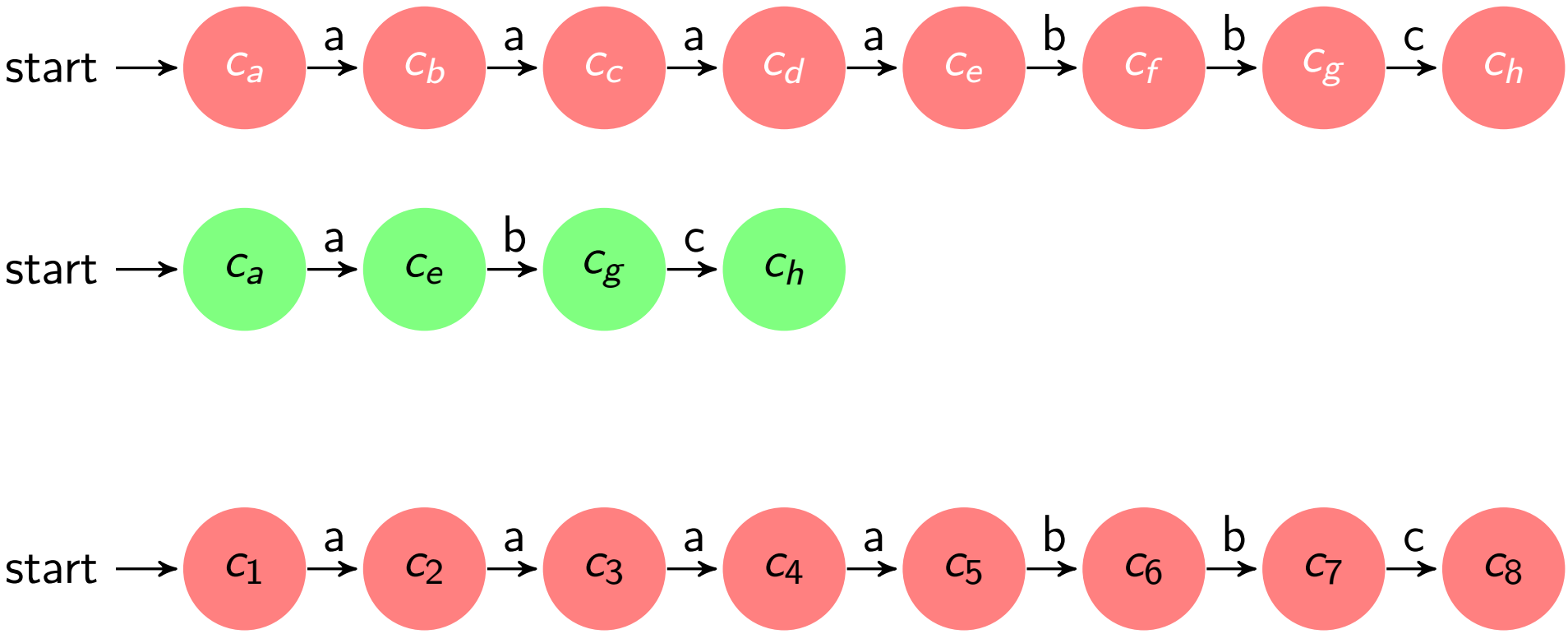
Ordered Pairs

A relation, R_e defined on the edges of $\mathcal{W}_1$ and $\mathcal{W}_2$ is given by $(e_1, e_2) \in R_e, e_1 \in E_1$ and $e_2 \in E_2$ where, $L_e(e_1) = L_e(e_2)$.

Predecessor

The subset of ordered pairs, Predecessor $Pred^r(Y)$ is defined from the set of ordered pairs, $(e_1, e_2) \in R_e$ represented by the Y is:
$Pred^r(Y) = \{(e_1, e_2) \in Y \mid \forall e_1', e_1 \rightarrowtail$
e_1'implies $\exists$ an il-path fragment $e_2 \rightarrowtail \ldots \rightarrowtail e_{m,2} \rightarrowtail e_2', \forall i \leq$
$m, (e_1, e_{i,2}) \in Y \wedge (e_1', e_2') \in Y,$ and $\forall e_2', e_2 \rightarrowtail e_2'$ implies $\exists$ an ilpath-fragment $e_1 \rightarrowtail \ldots \rightarrowtail e_{m,1} \rightarrowtail e_1', \forall i \leq m, (e_{i,1}, e_2) \in Y \wedge (e_1', e_2') \in Y\}$.


start
c_a
a
c_b
a
c_c
a
c_d
a
c_e
b
c_f
b
c_g
c
c_h
start
c_a
a
c_e
b
c_g
c
c_h


start
c_a
a
c_b
a
c_c
a
c_d
a
c_e
b
c_f
b
c_g
c
c_h
start
c_a
a
c_e
b
c_g
c
c_h
start
c_1
a
c_2
a
c_3
a
c_4
a
c_5
b
c_6
b
c_7
c
c_8

Model Assumptions

- Comparing LPSs is to be compared on the same state space (edge labels) as required by KLD.
- There is no self loop on the states.

Computing:Greatest Fixed Point

Input: Set of Ordered Pairs, R_e

Output: Set of ordered pairs in the greatest fixed point, Y_∞.

1. $Y := R_e$;
2. $Y' := 0$;
3. $H(\mathcal{W}_1 \| \mathcal{W}_2) = 0$;
4. while $(Y \neq Y')$
5. $\{$
6. $Y' := Y$;
7. $Y := Y \cap Pred^r(Y)$;
8. $H(\mathcal{W}_1 \| \mathcal{W}_2) = H(\mathcal{W}_1 \| \mathcal{W}_2) + P_c(Fst(Y), Pre(Fst(Y)) log(\frac{P_c(Fst(Y), Pre(Fst(Y)))}{P_c(Snd(Y), Pre(Snd(Y)))}$
9. $\}$
10. $Y_\infty = Y'$

Termination of the algorithm

Lemma

The algorithm terminates after finite number of steps and computes fixed point, given by $Y = Pred^r(Y)$.

Proof sketch:
Finite number of ordered pairs of edges in R_e.
The algorithm computes the fixed point, i.e $Y = Pred^r(Y)$.

Complexity of the algorithm

The time complexity of the algorithm is $O(m^2)$ where $m = \mid R_e \mid$. In the worst case, the set of ordered pairs in $Pred^{st}(Y)$ is constructed by removing a pair (e_1, e_2) at a time . The while loop iterates m times over m computations in $Pred^{r}(Y)$.

Correctness

Lemma

If $e_1 \equiv_r^{i+1} e_2$ *then* $e_1 \equiv_r^{i} e_2$.

Lemma

If $(e_1, e_2) \in Y_{i+1}$ *then* $e_1 \equiv_r^{i+1} e_2$.

Lemma

If $e_1 \equiv_r^{i+1} e_2$ *then* $(e_1, e_2) \in Y_{i+1}$.

Quantification of Errors

Approximation leads to *errors*. What are the potential errors?

Quantification of Errors

Approximation leads to *errors*. What are the potential errors? Can we quantify it?

- A path segment has *il*-path and the other does not, $AP - Error$
- Both the paths have *il* path, $CP - Error$.

AP-Error and CP-Error

1. AutoPath(AP)-Error: A trace has compact probability and the other trace does not. The error is given by
   Use Compact Probability of *il*-path fragment and maximum probability, p_{max} of probabilities on the edges in the *il* path :
   ApError= $p_{max} log \frac{p_{max}}{P_c(p_1,...,p_k)}$
2. CoPath(CP)-Error: Both the paths have *il* path fragments.
   Sum the error probabilities on the edges in the *il* path :

$$\text{CpError} = \sum_i cError_i$$

where $cError_i = p^i_{max} log \frac{p^i_{max}}{P_c(p^i_1,...,p^i_k)}$ where $i = 1, 2$.

Current Directions

Modeling

- A notion of similarity among probabilitic models is created.
- Ongoing work on the statistics based on the errors such that models can be generated.

- Thank You
- Thank You to Reviewers
- Thank You to Organizers