Summarizing the Solution Space in Tumor Phylogeny Inference by Multiple Consensus Trees

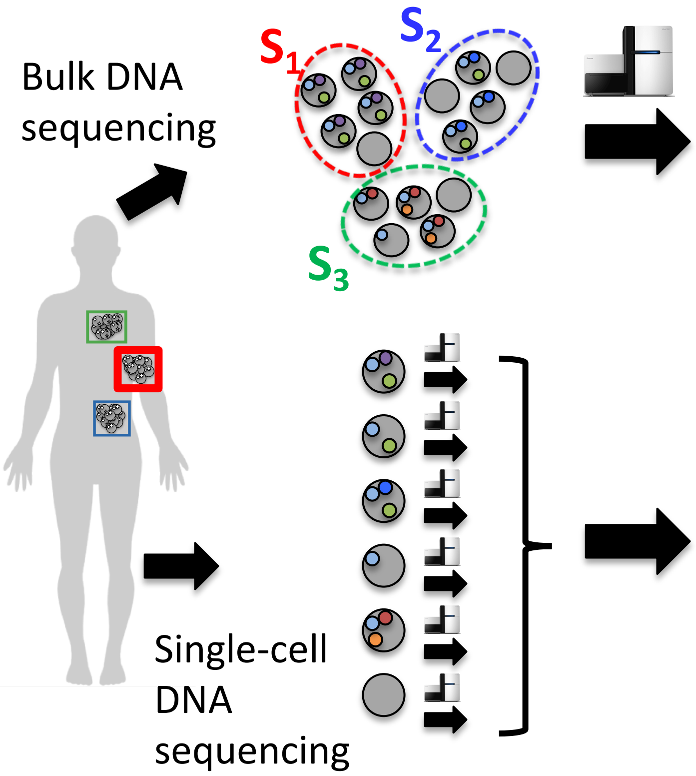
Nuraini Aguse*, Yuanyuan Qi* and Mohammed El-Kebir
University of Illinois at Urbana Champaign, Department of Computer Science

RECOMB-CCB 2019

*Joint first authorship

Accepted at ISMB/ECCB 2019

Additional Challenge in Cancer Phylogenetics



n **mutations**

$$
\begin{array}{c} \\ \\ m \text{ samples} \end{array}
\begin{array}{c} S_1 \\ S_2 \\ S_3 \end{array}
\begin{pmatrix}
0.8 & 0.8 & 0.8 & 0.0 & 0.0 & 0.0 \\
0.7 & 0.6 & 0.0 & 0.6 & 0.0 & 0.0 \\
0.8 & 0.0 & 0.0 & 0.0 & 0.6 & 0.4
\end{pmatrix}
$$

n **mutations**

$$
m \text{ cells}
\begin{pmatrix}
1 & 1 & 1 & 0 & 0 & \color{red}{1} \\
1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & \color{red}{0} & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & \color{red}{?} & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}
$$

Incorrect entries:

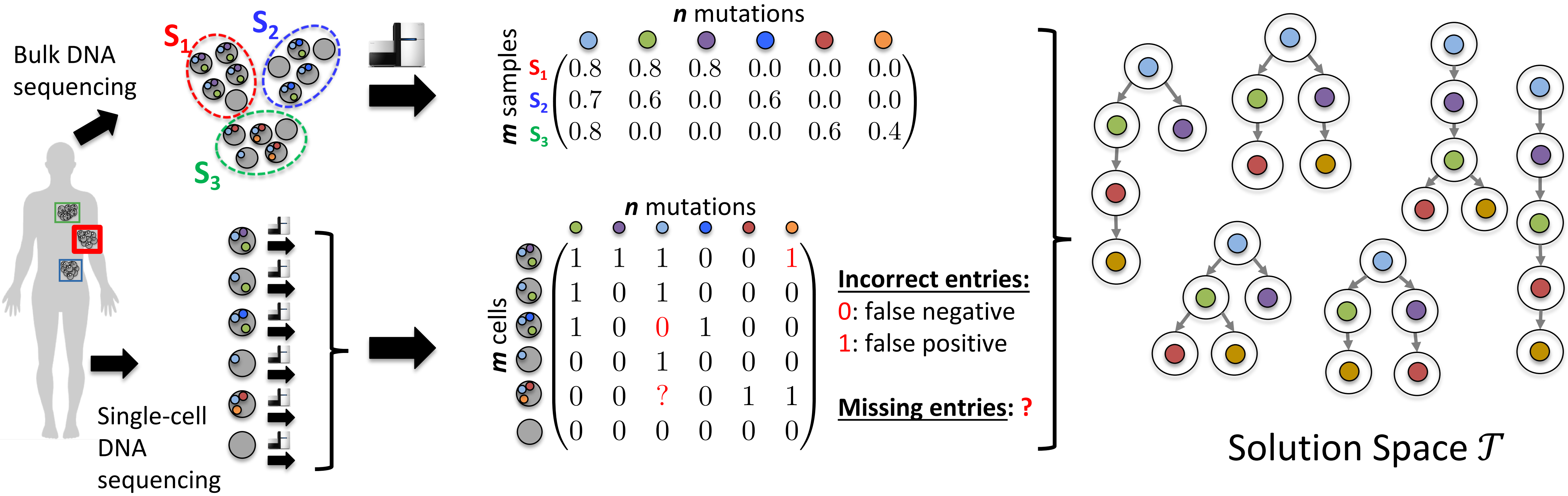
0: false negative

1: false positive

Missing entries: ?

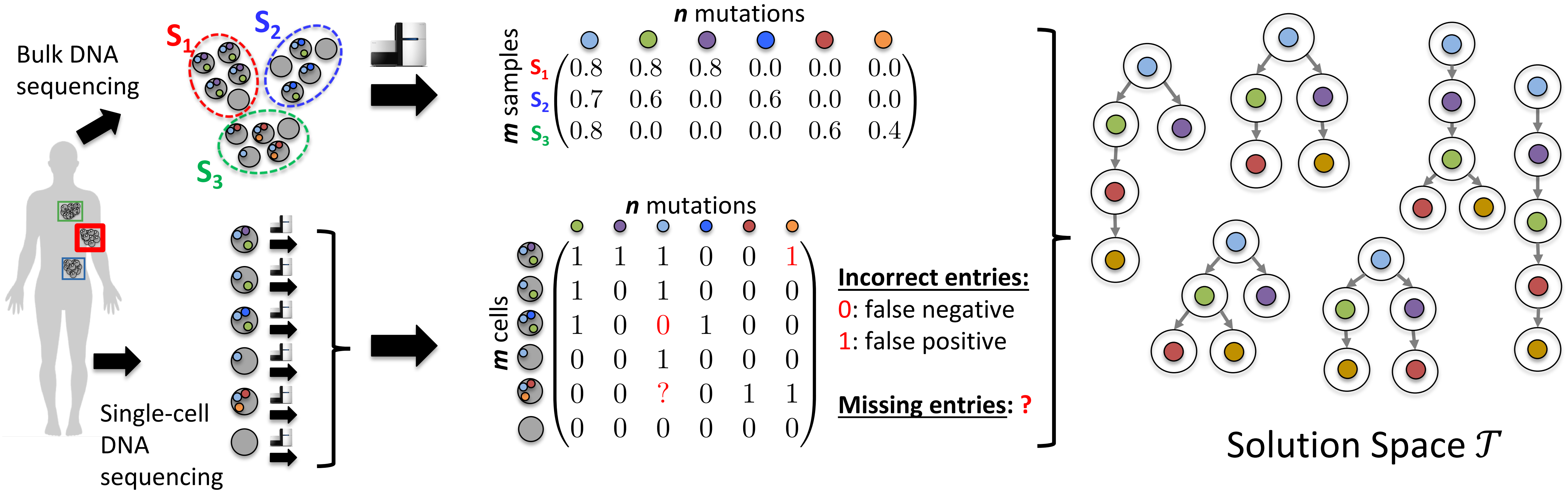
Phylogeny inference from mixtures of/incomplete measurements of leaves

Additional Challenge in Cancer Phylogenetics



Phylogeny inference from mixtures of/incomplete measurements of leaves	Non-uniqueness of solutions: alternative solutions with varying leaf sets

Additional Challenge in Cancer Phylogenetics



Phylogeny inference from mixtures of/incomplete measurements of leaves

Non-uniqueness of solutions: alternative solutions with varying leaf sets

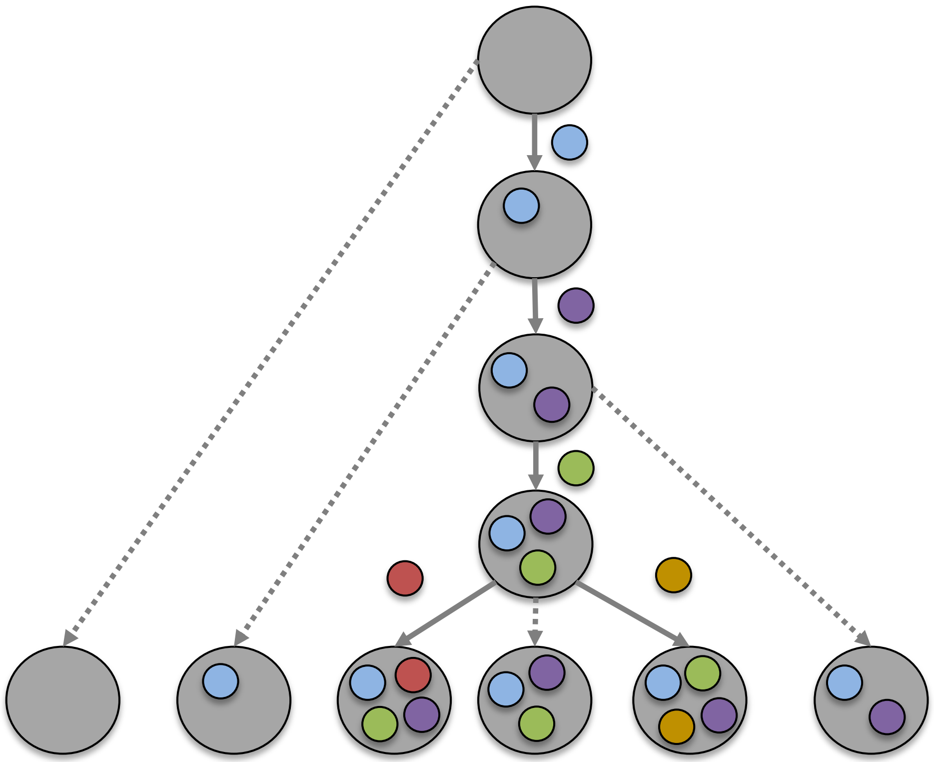
Question: How to **summarize solution space $\mathcal{T}$** in order to remove inference errors and identify dependencies among mutations?

Outline

- Problem Statement
  - Previous work
  - Problem statement
  - Combinatorial characterization of solutions
  - Complexity

- Method & Results
  - Exact algorithm
  - Heuristic algorithm
  - Model selection

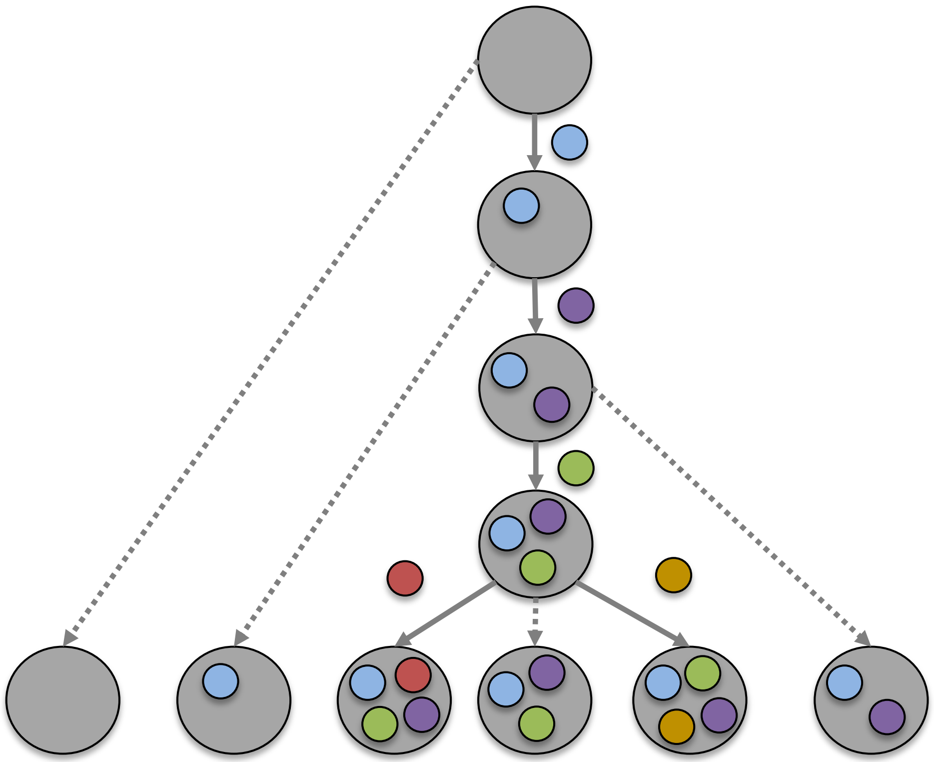
Phylogenetic Trees vs. Mutation Trees



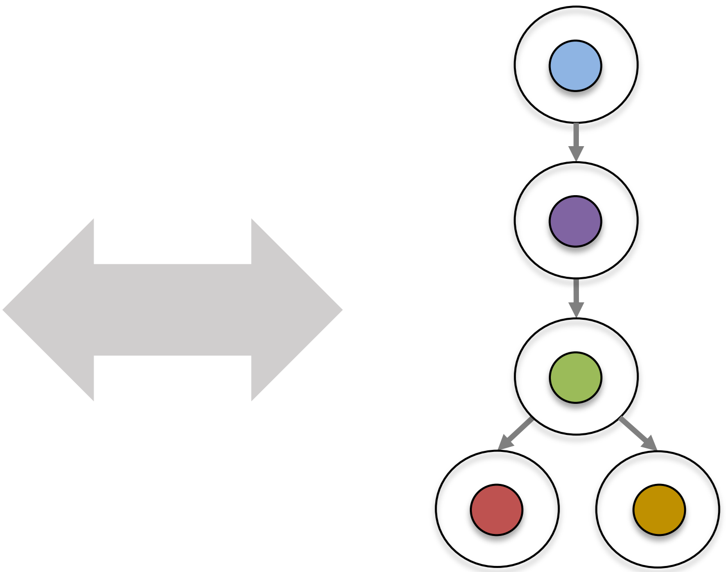
Phylogenetic Tree

Infinite sites assumption (ISA): each mutation is introduced once and never subsequently lost

Phylogenetic Trees vs. Mutation Trees



Phylogenetic Tree



Mutation Tree

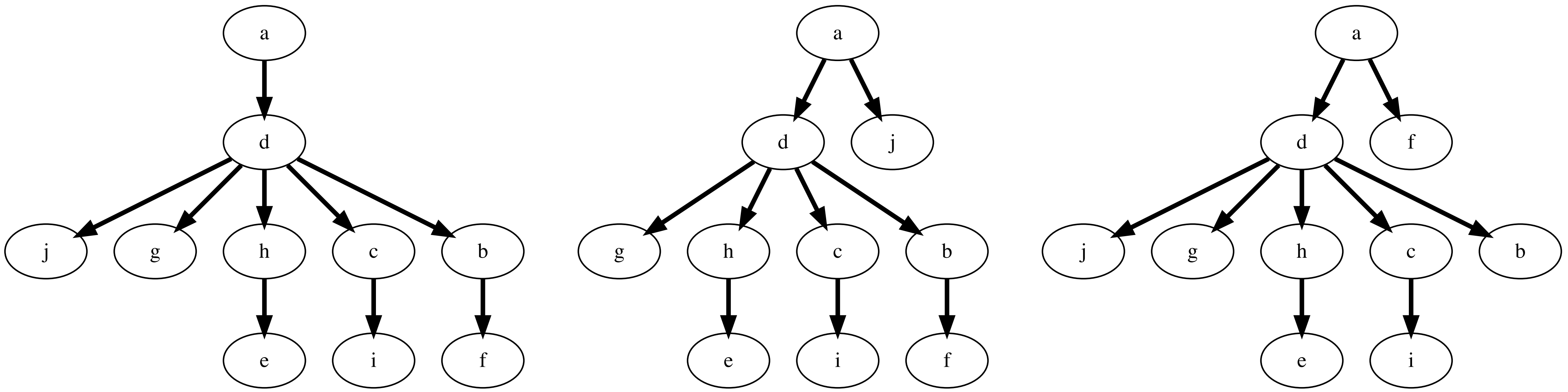
Infinite sites assumption (ISA): each mutation is introduced once and never subsequently lost

Under ISA, a phylogenetic tree may be equivalently* represented by a mutation tree

Solution Space of Lung Cancer Patient CRUK0037

Jamal-Hanjani et al. (2017). *New England Journal of Medicine*, *376*(22), 2109–2121.

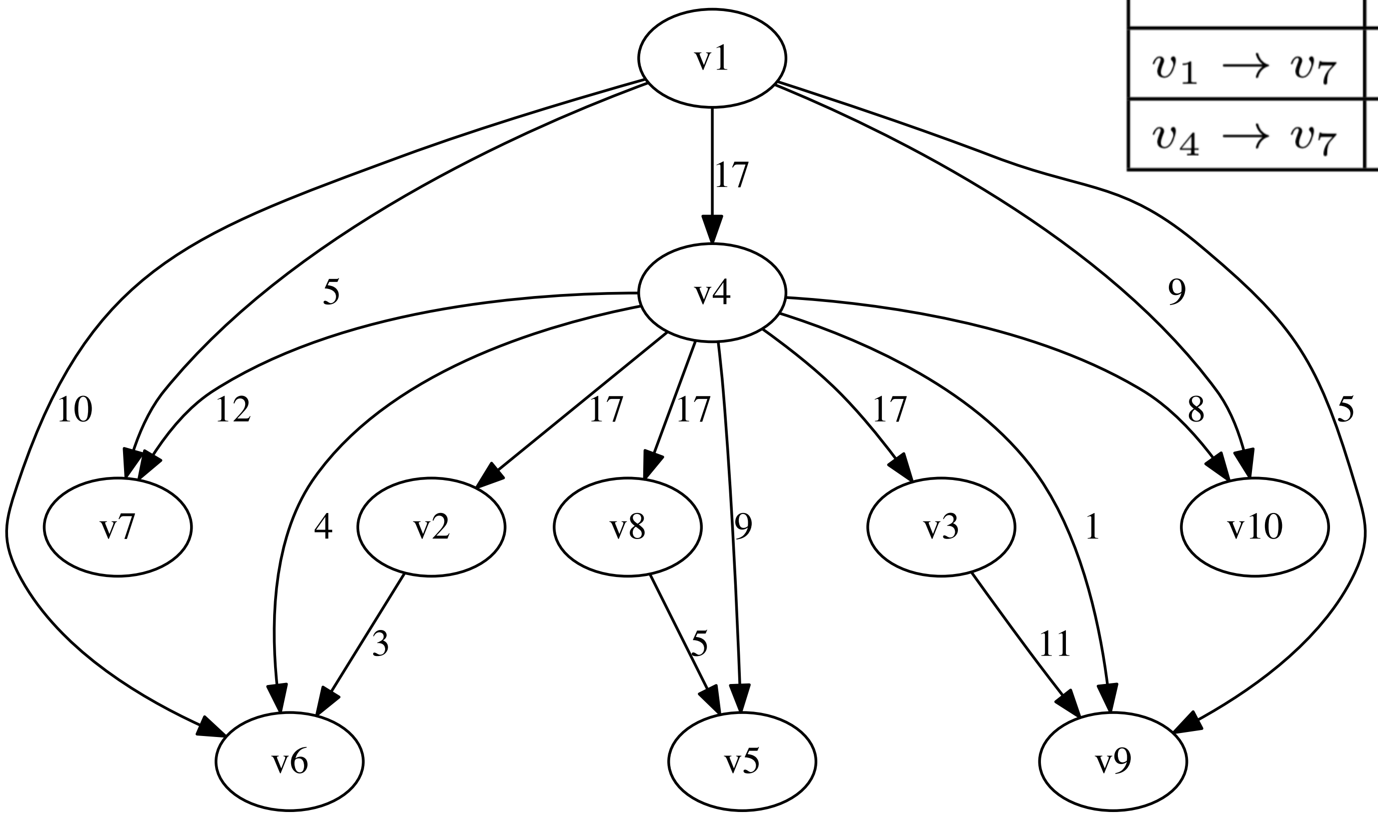
Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037



...
[14 more]

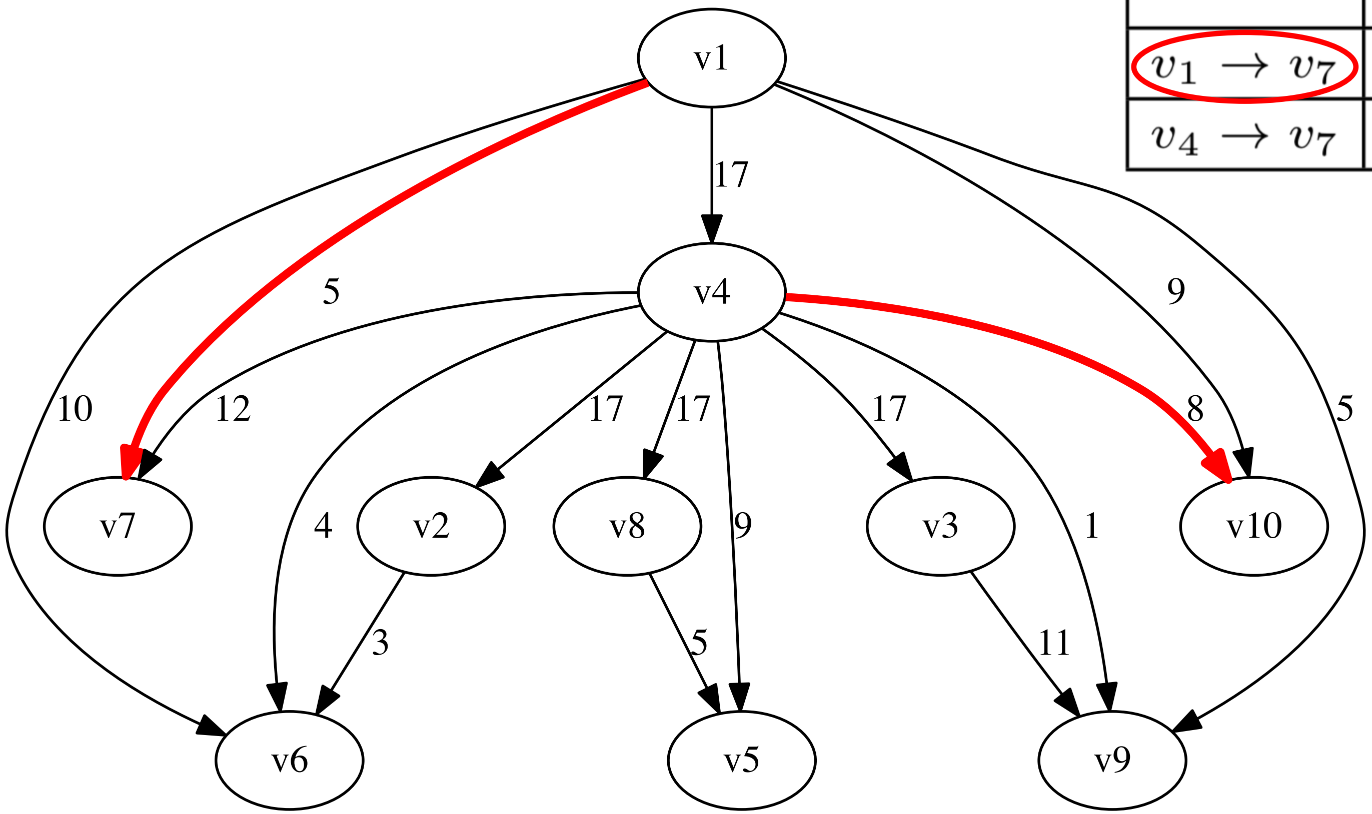
Question: How to **summarize solution space** in order to remove inference errors and identify dependencies among mutations?

Parent-child Graph: Union of all Edges in $\mathcal{T}$



	$v_4 \to v_5$		$v_8 \to v_5$	
	$v_1 \to v_{10}$	$v_4 \to v_{10}$	$v_1 \to v_{10}$	$v_4 \to v_{10}$
$v_1 \to v_7$	2	0	3 (d)	0
$v_4 \to v_7$	2 (b)	5 (e)	2	3

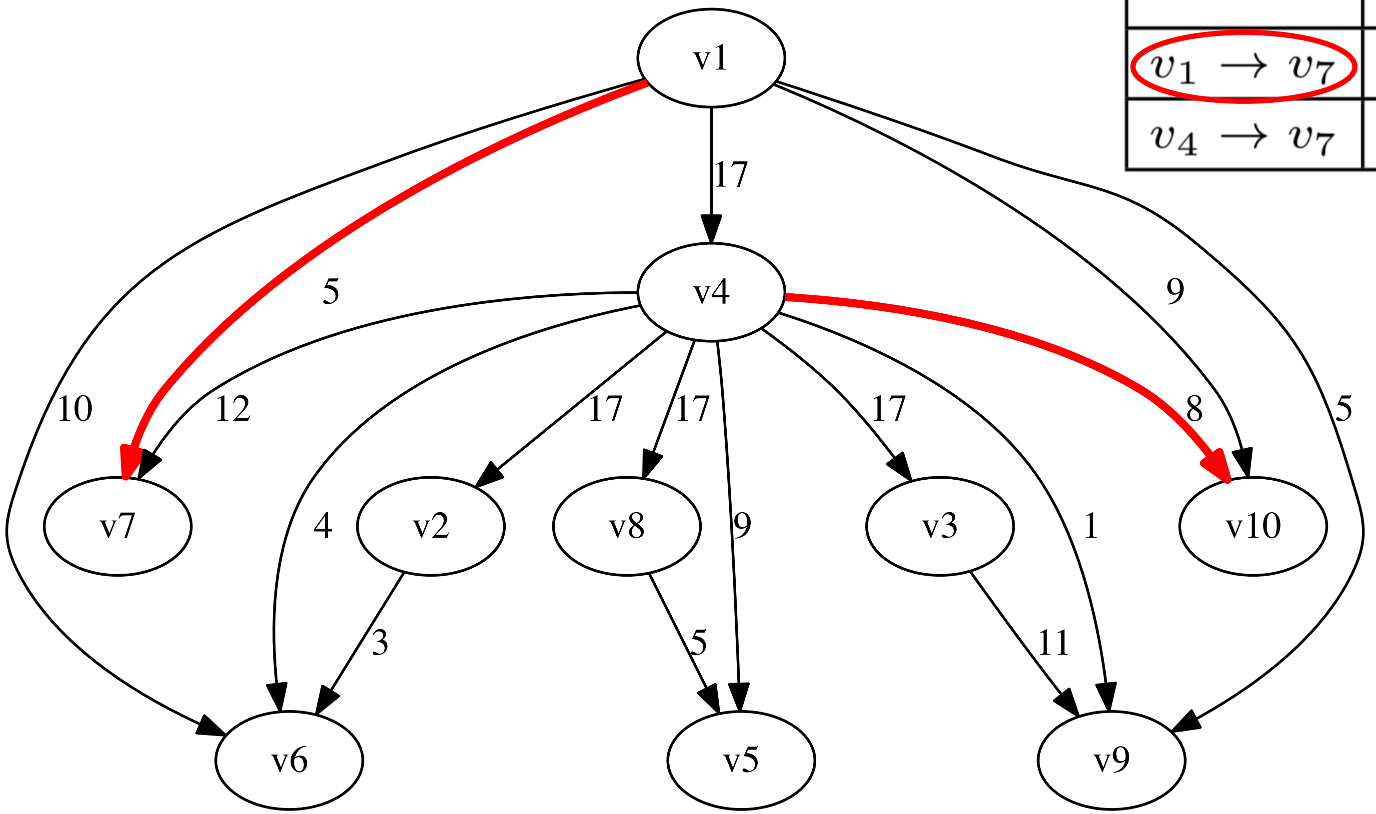
Parent-child Graph: Union of all Edges in $\mathcal{T}$



	$v_4 \to v_5$		$v_8 \to v_5$	
	$v_1 \to v_{10}$	$v_4 \to v_{10}$	$v_1 \to v_{10}$	$v_4 \to v_{10}$
$v_1 \to v_7$	2	0	3 (d)	0
$v_4 \to v_7$	2 (b)	5 (e)	2	3

The parent-child graph does not capture patterns of mutual exclusivity

Parent-child Graph: Union of all Edges in $\mathcal{T}$



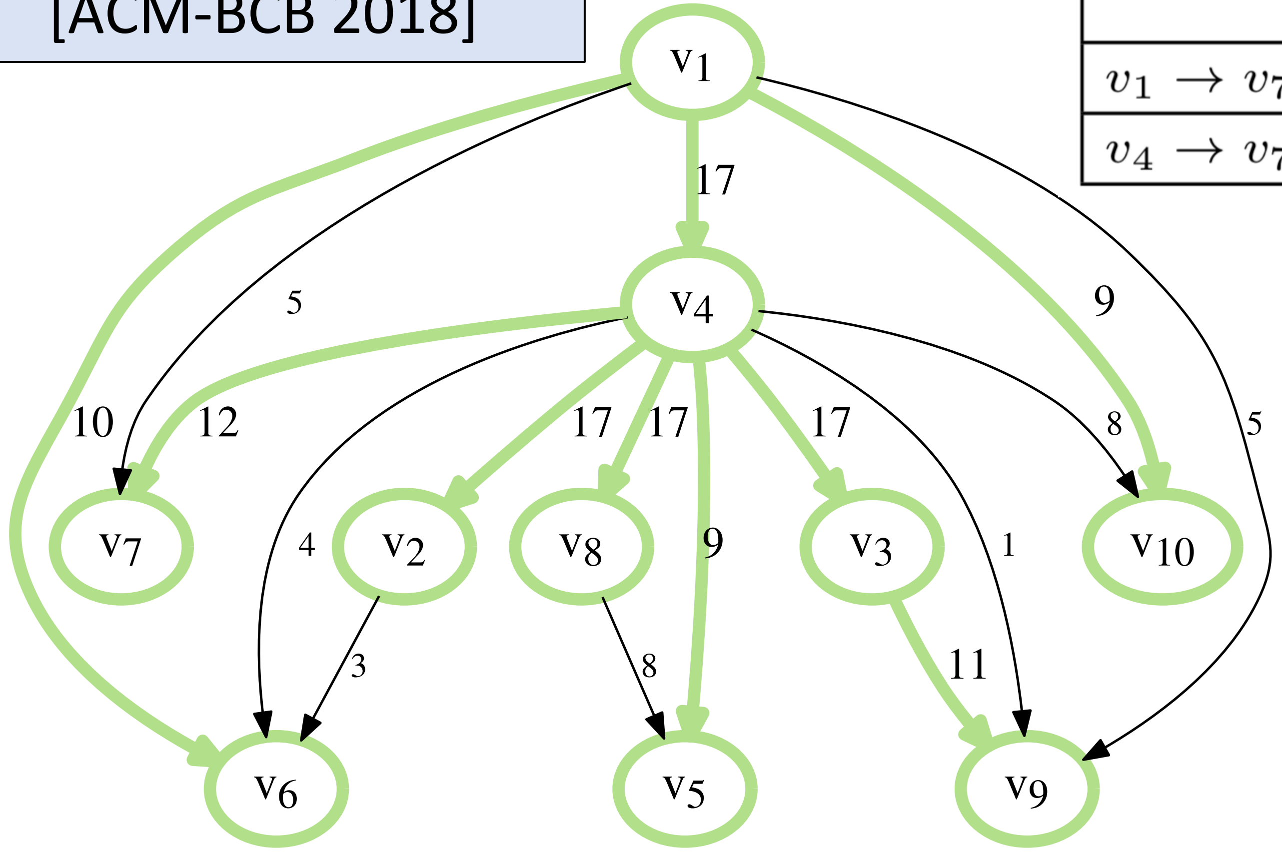
	$v_4 \to v_5$		$v_8 \to v_5$	
	$v_1 \to v_{10}$	$v_4 \to v_{10}$	$v_1 \to v_{10}$	$v_4 \to v_{10}$
$v_1 \to v_7$	2	0	3 (d)	0
$v_4 \to v_7$	2 (b)	5 (e)	2	3

The parent-child graph does not capture patterns of mutual exclusivity

Question: Can we infer a single consensus tree?

Single Consensus Tree: Max Weight Spanning Tree

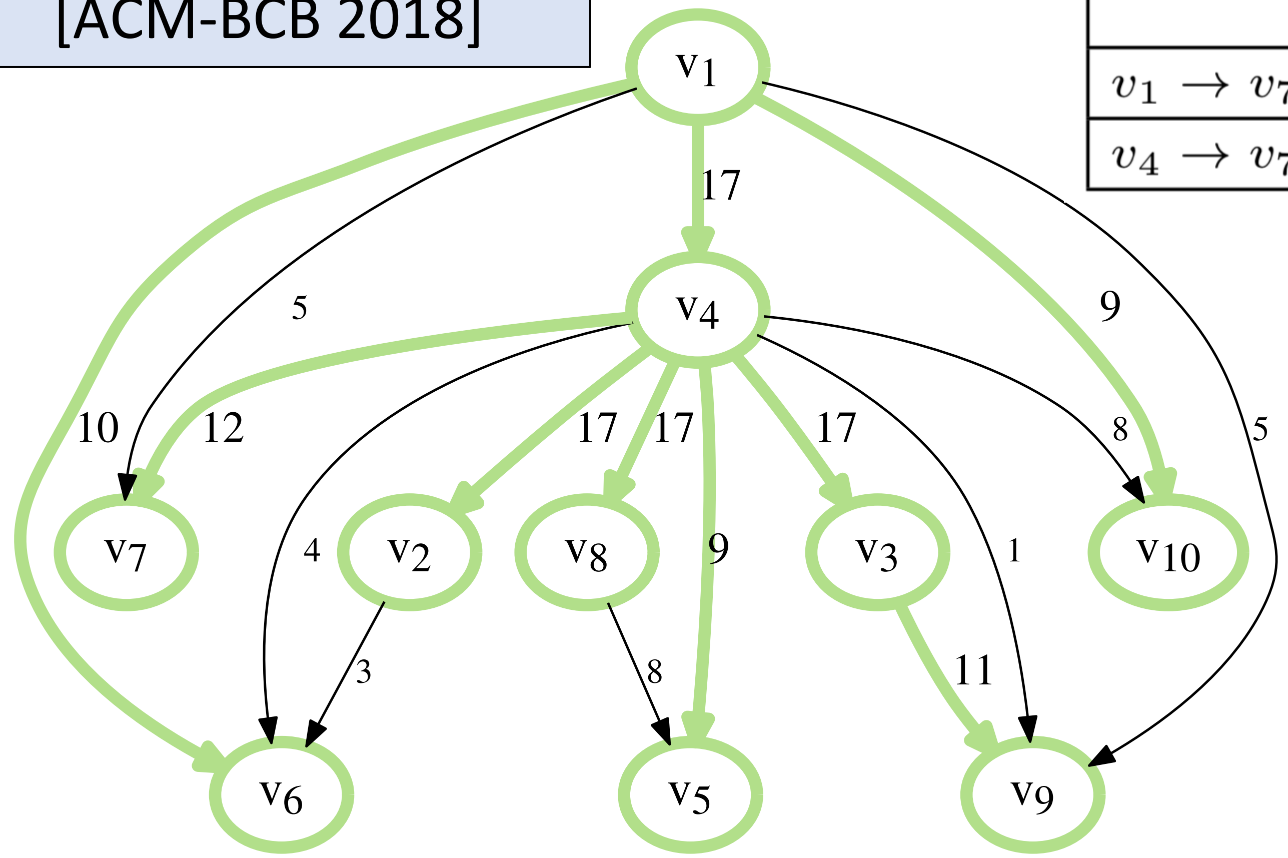
Oesper and colleagues. [ACM-BCB 2018]



	$v_4 \rightarrow v_5$		$v_8 \rightarrow v_5$	
	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$	$v_1 \rightarrow v_{10}$	$v_4 \rightarrow v_{10}$
$v_1 \rightarrow v_7$	2	0	3 (d)	0
$v_4 \rightarrow v_7$	2 (b)	5 (e)	2	3

Single Consensus Tree: Max Weight Spanning Tree

Oesper and colleagues.
[ACM-BCB 2018]



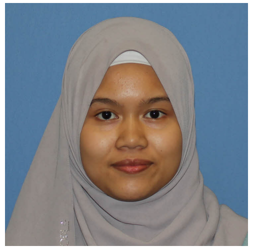
	$v_4 \to v_5$		$v_8 \to v_5$	
	$v_1 \to v_{10}$	$v_4 \to v_{10}$	$v_1 \to v_{10}$	$v_4 \to v_{10}$
$v_1 \to v_7$	2	0	3 (d)	0
$v_4 \to v_7$	2 (b)	5 (e)	2	3

Inaccurate summary for diverse solution spaces

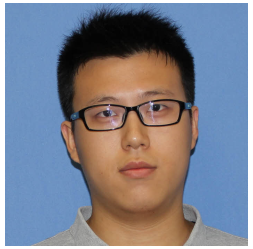
Question: How about inferring multiple consensus trees?

Multiple Consensus Trees Problem

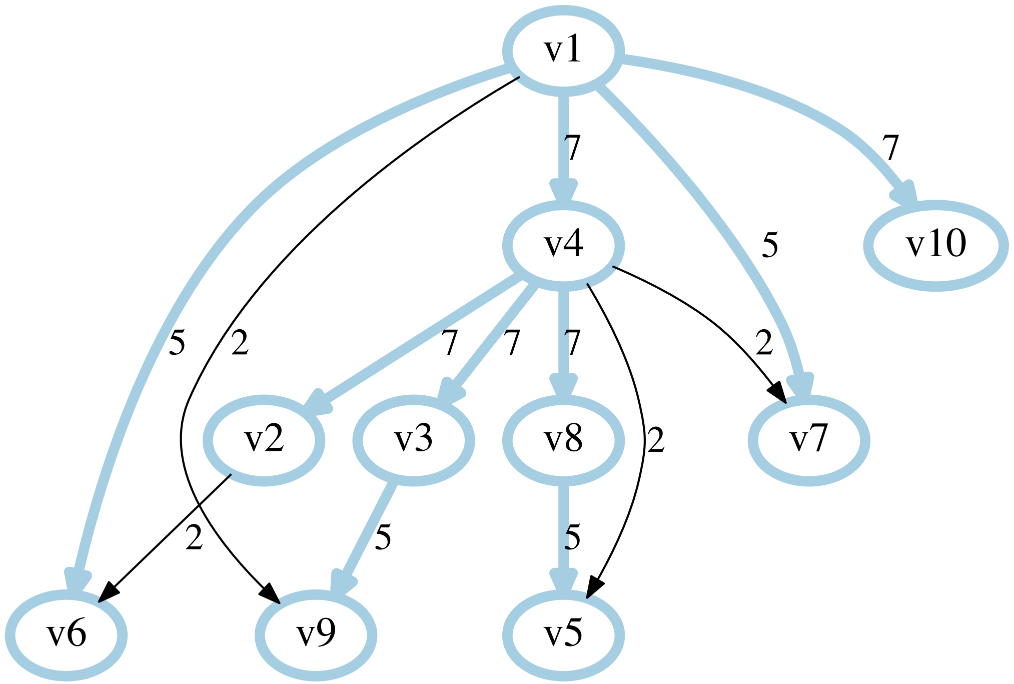
Simultaneous clustering and consensus tree inference

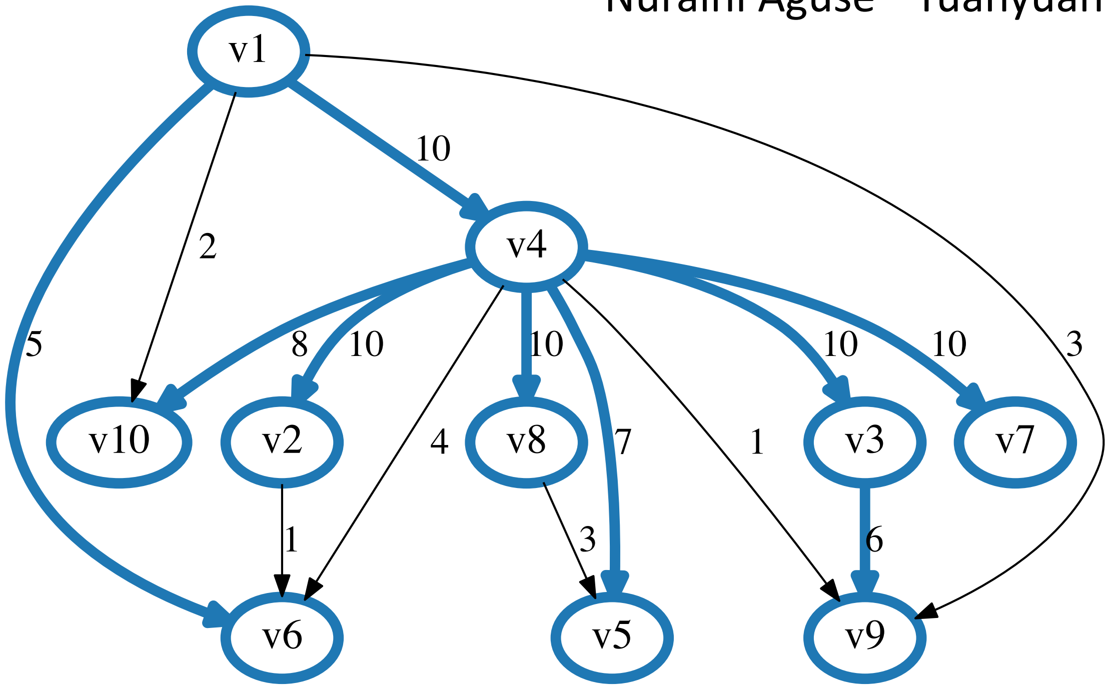


Nuraini Aguse

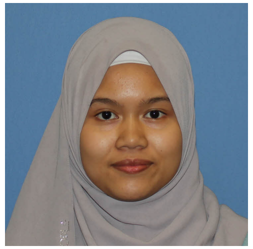
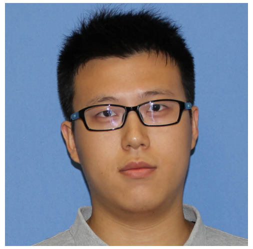


Yuanyuan Qi



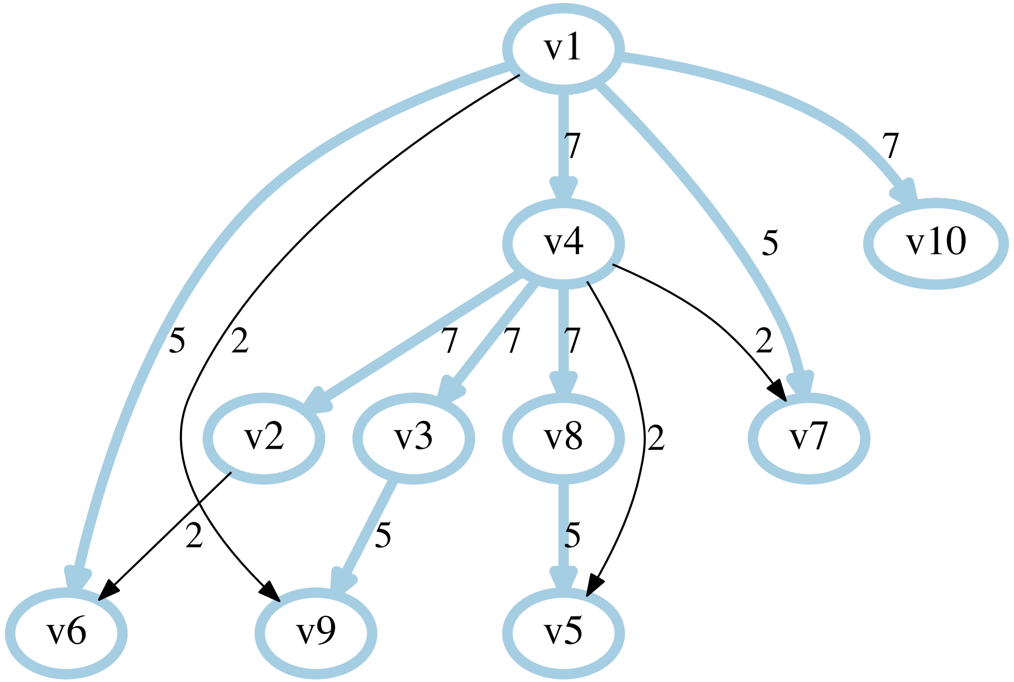


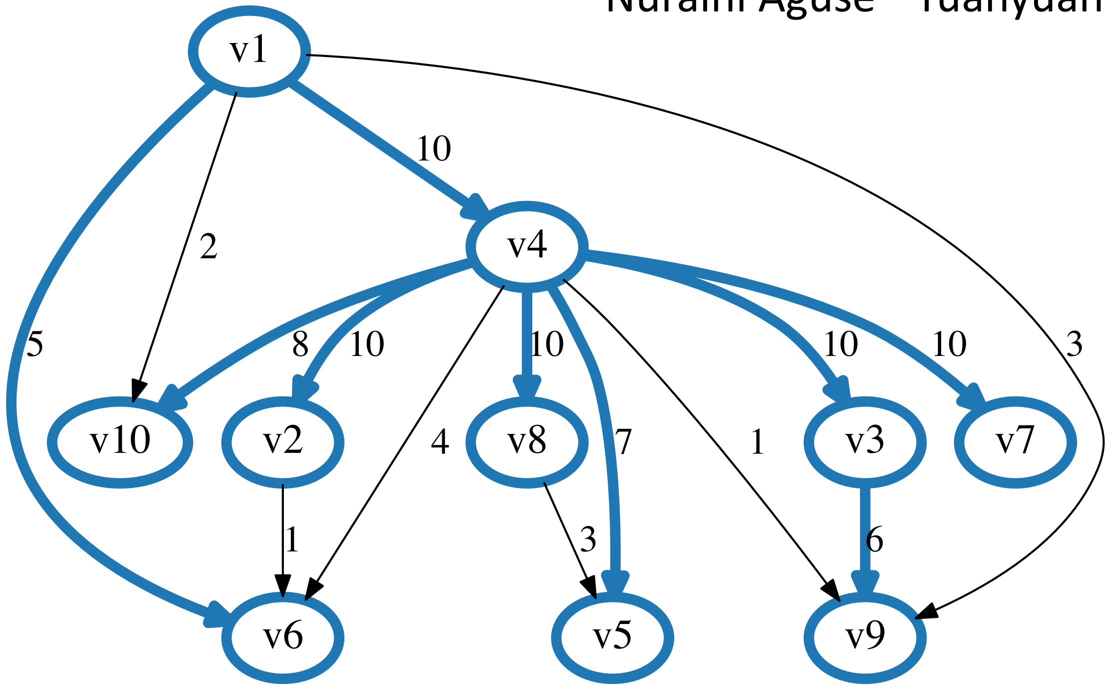
Multiple Consensus Trees Problem

Nuraini Aguse Yuanyuan Qi

Simultaneous clustering and consensus tree inference

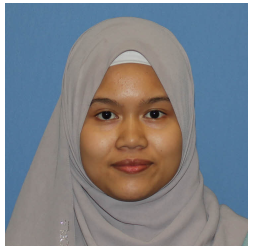
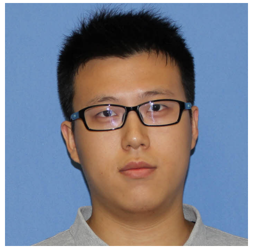




Multiple Consensus Trees (MCT): [ISMB/ECCB 2019]

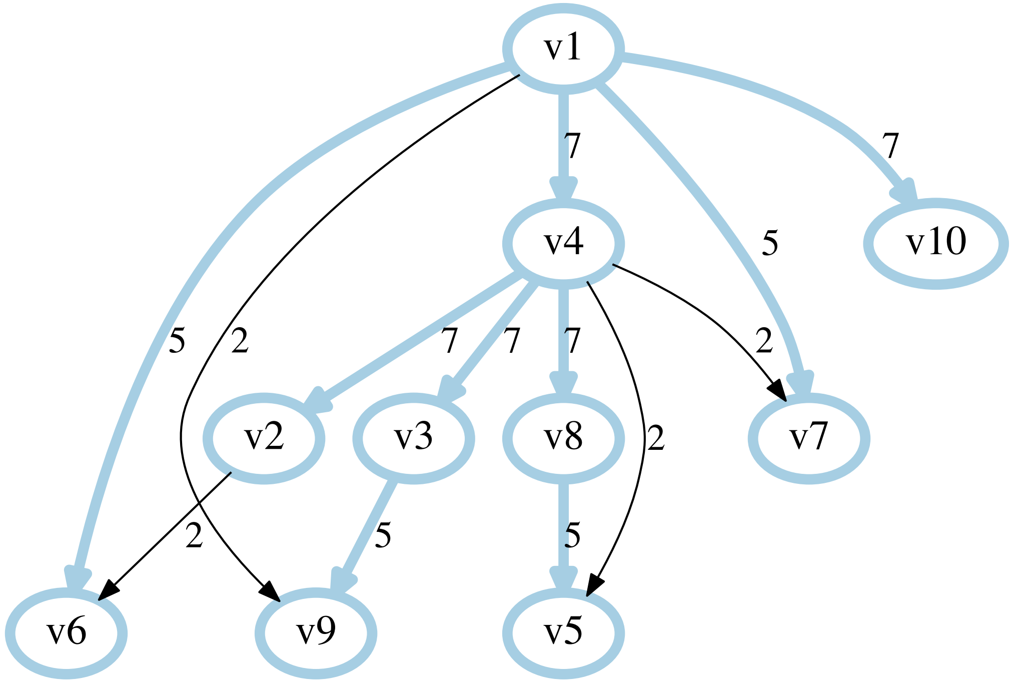
Given trees $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum

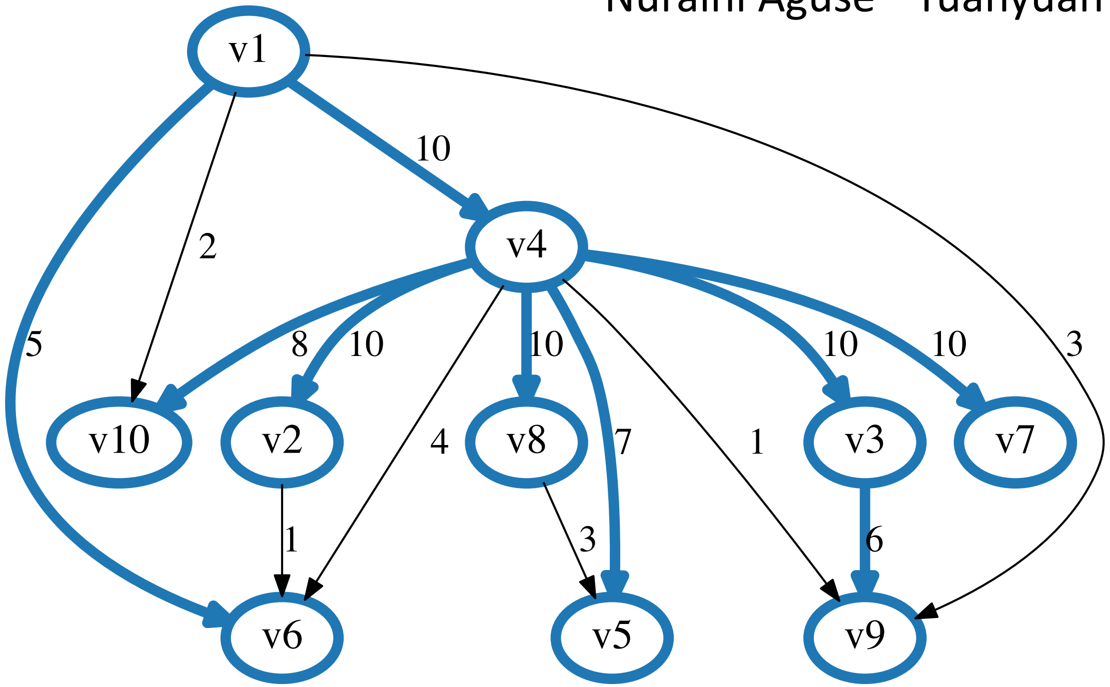
Multiple Consensus Trees Problem

Nuraini Aguse Yuanyuan Qi

Simultaneous clustering and consensus tree inference

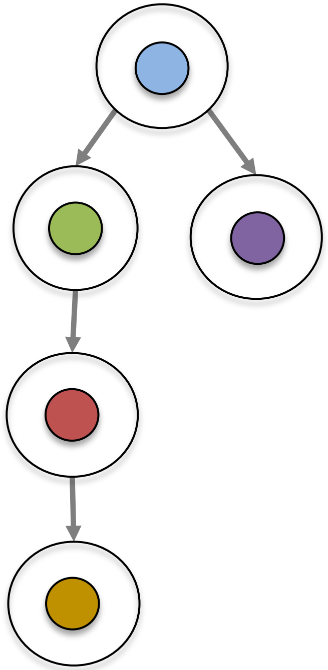




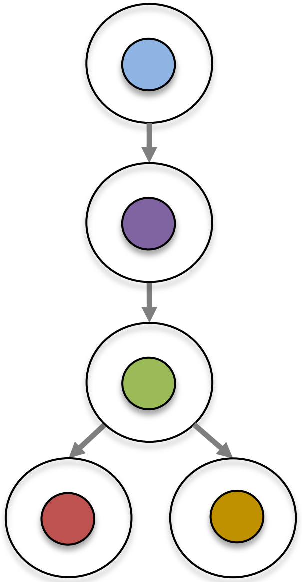
Multiple Consensus Trees (MCT): [ISMB/ECCB 2019]

Given trees $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering σ : $[n] \rightarrow [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^{n} \boxed{d(T_i, R_{\sigma(i)})}$ is minimum

Parent-child Distance Function

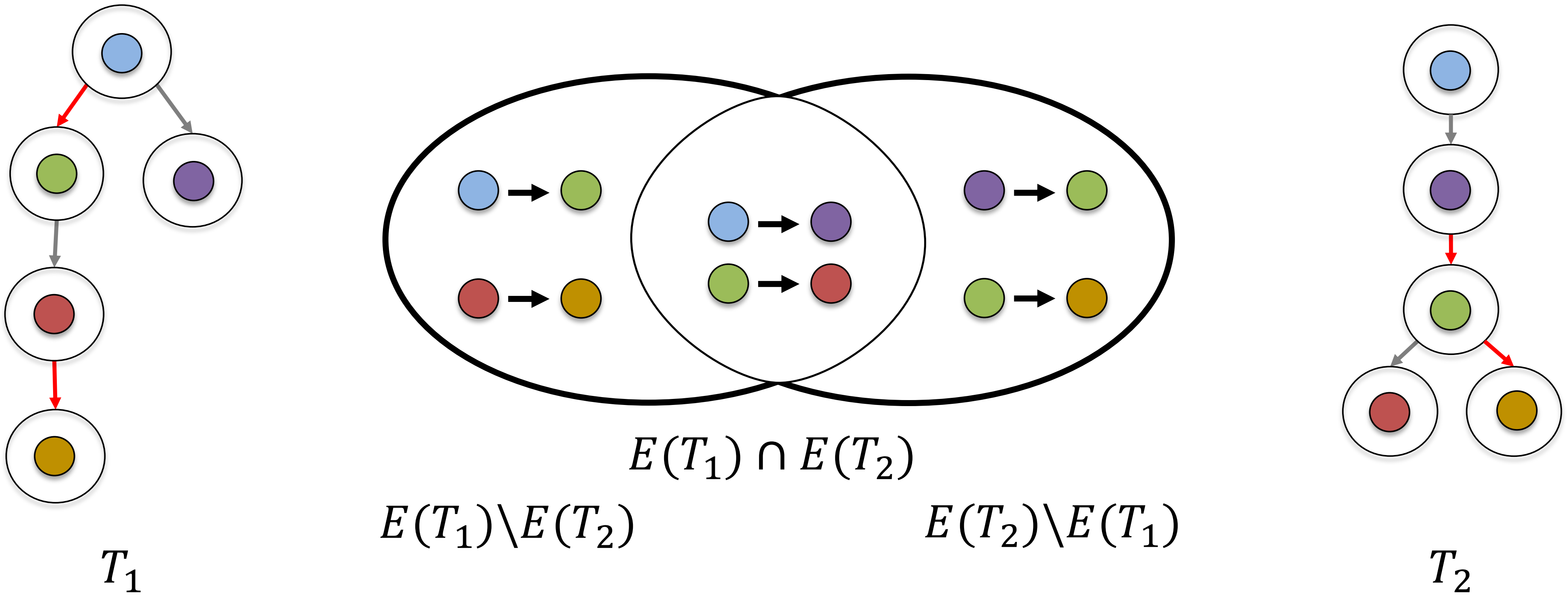


T_1

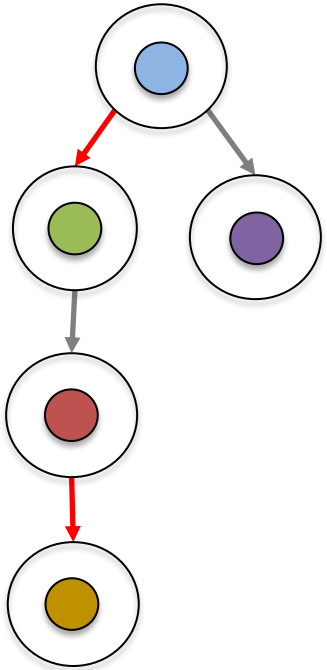


T_2

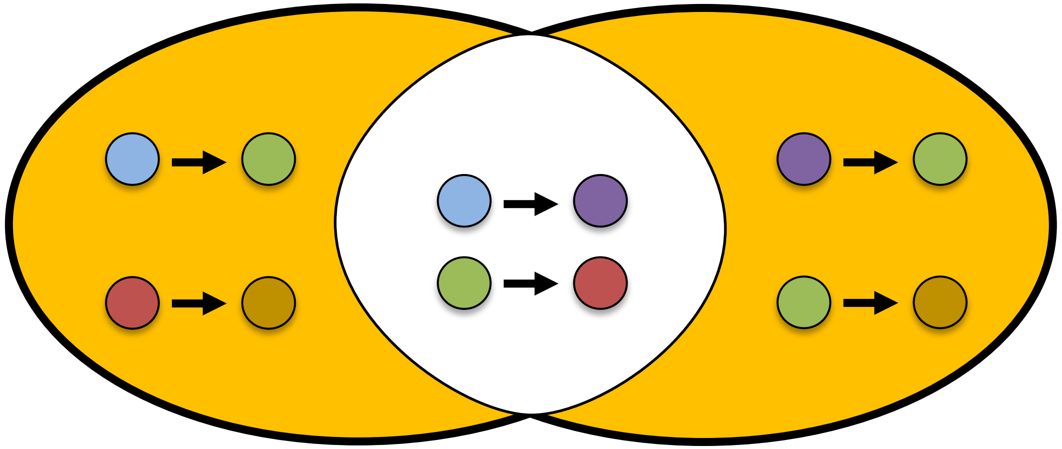
Parent-child Distance Function



Parent-child Distance Function

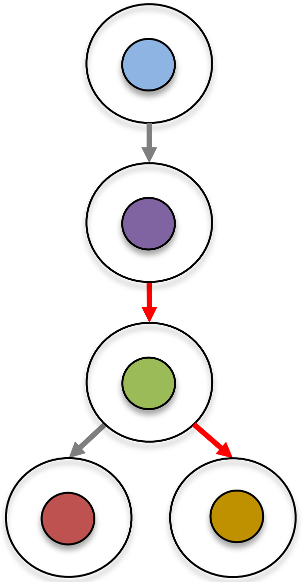


T_1



$E(T_1) \cap E(T_2)$

$E(T_1) \setminus E(T_2)$ $E(T_2) \setminus E(T_1)$



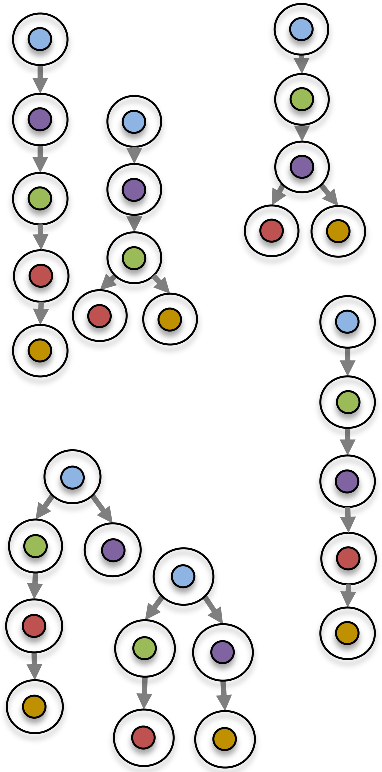
T_2

Parent-child distance $d(T_1, T_2)$ is the size of the symmetric difference of the edge sets

Here, $d(T_1, T_2) = |E(T_1) \setminus E(T_2)| + |E(T_2) \setminus E(T_1)| = 4.$

Combinatorial Characterization of Solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

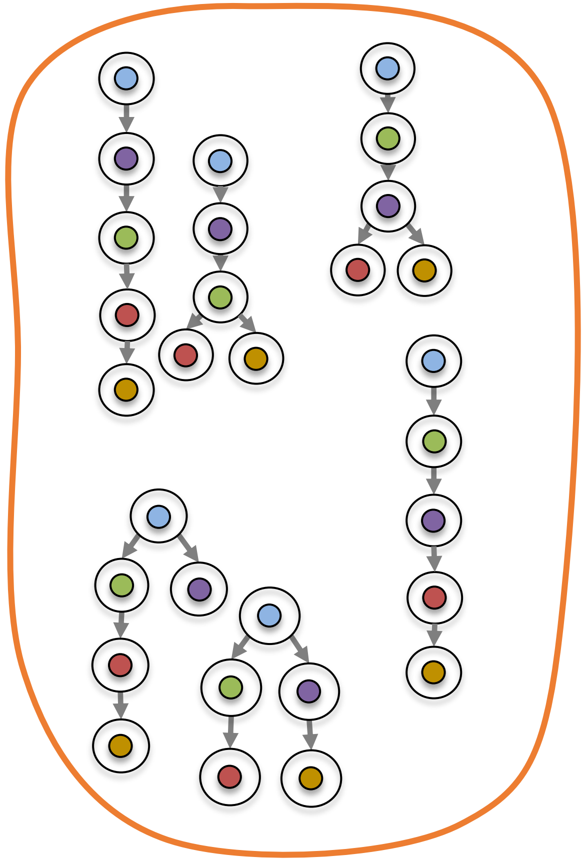


Solution Space $\mathcal{T}$

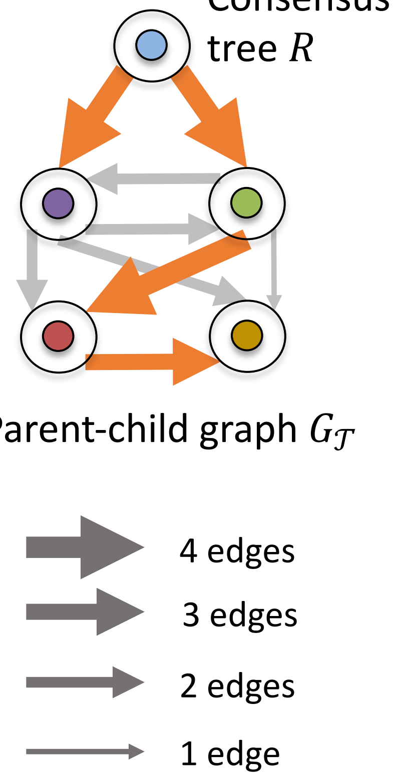
Combinatorial Characterization of Solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT



Solution Space $\mathcal{T}$

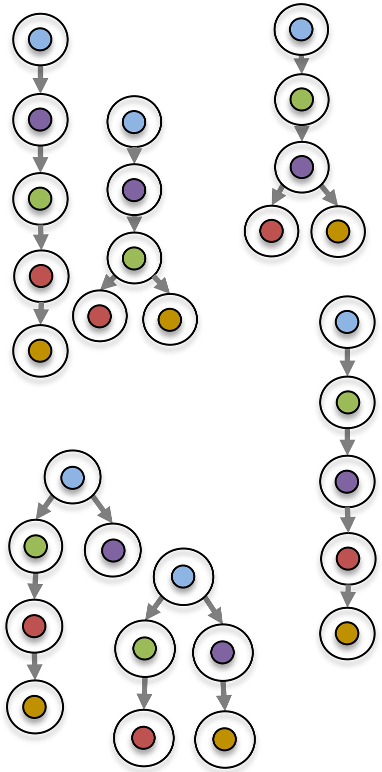


Combinatorial Characterization of Solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum



Solution Space $\mathcal{T}$

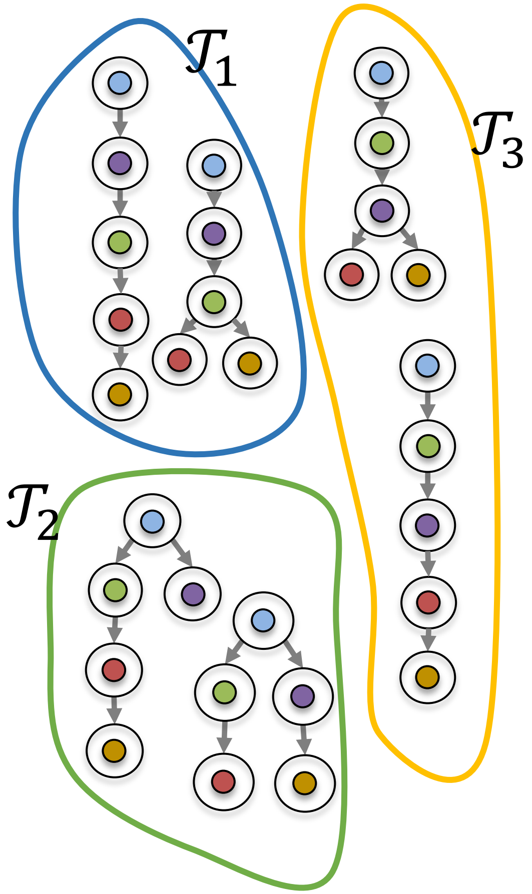
Combinatorial Characterization of Solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$, find consensus tree R s.t. $\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \to [k]$ and consensus trees $\mathcal{R} = \{R_1, \dots, R_k\}$ s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \to [k]$, MCT decomposes into k independent SCT instances



Solution Space $\mathcal{T}$

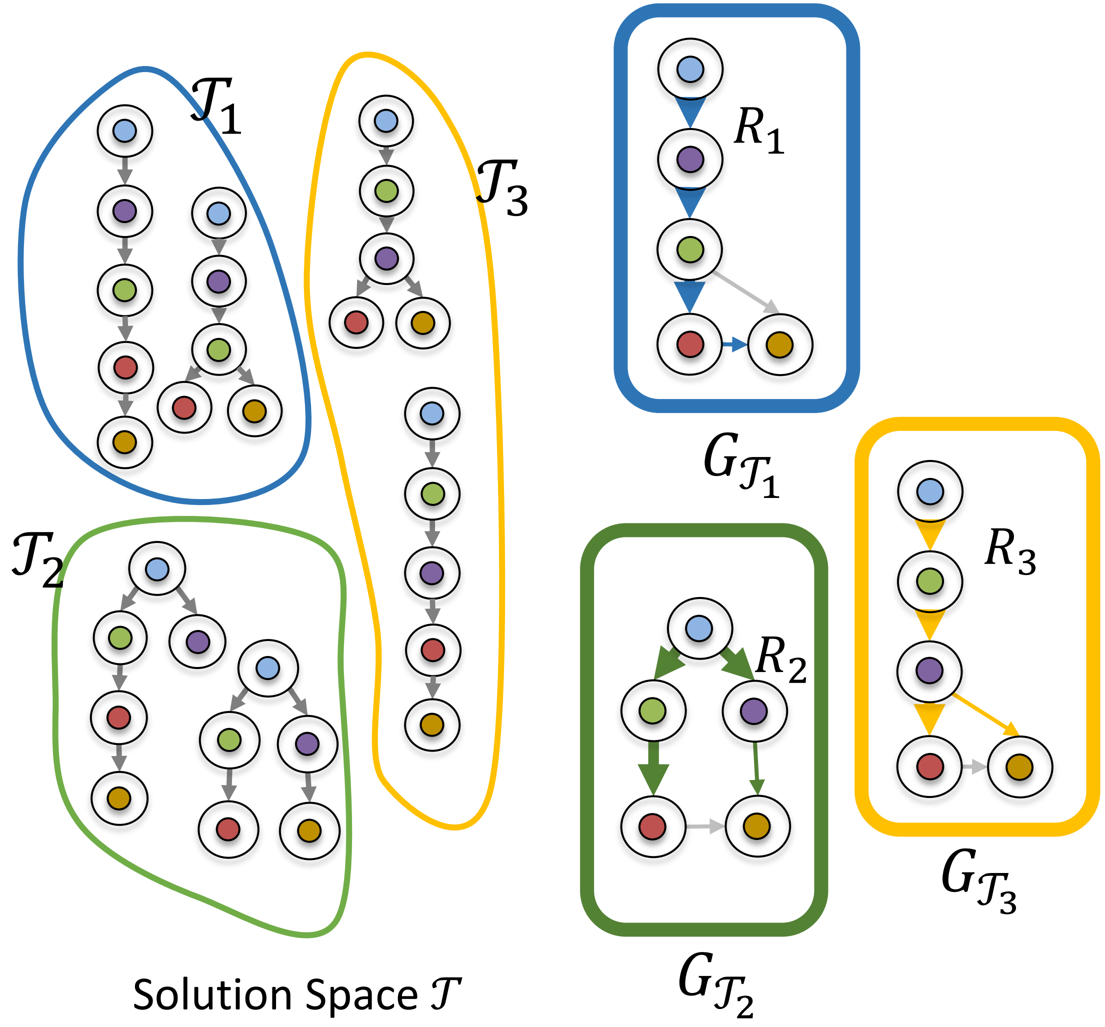
Combinatorial Characterization of Solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \to [k]$ ~~and consensus trees~~ $\cancel{\mathcal{R} = \{R_1, \ldots, R_k\}}$
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum
where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \to [k]$, MCT decomposes into
k independent SCT instances



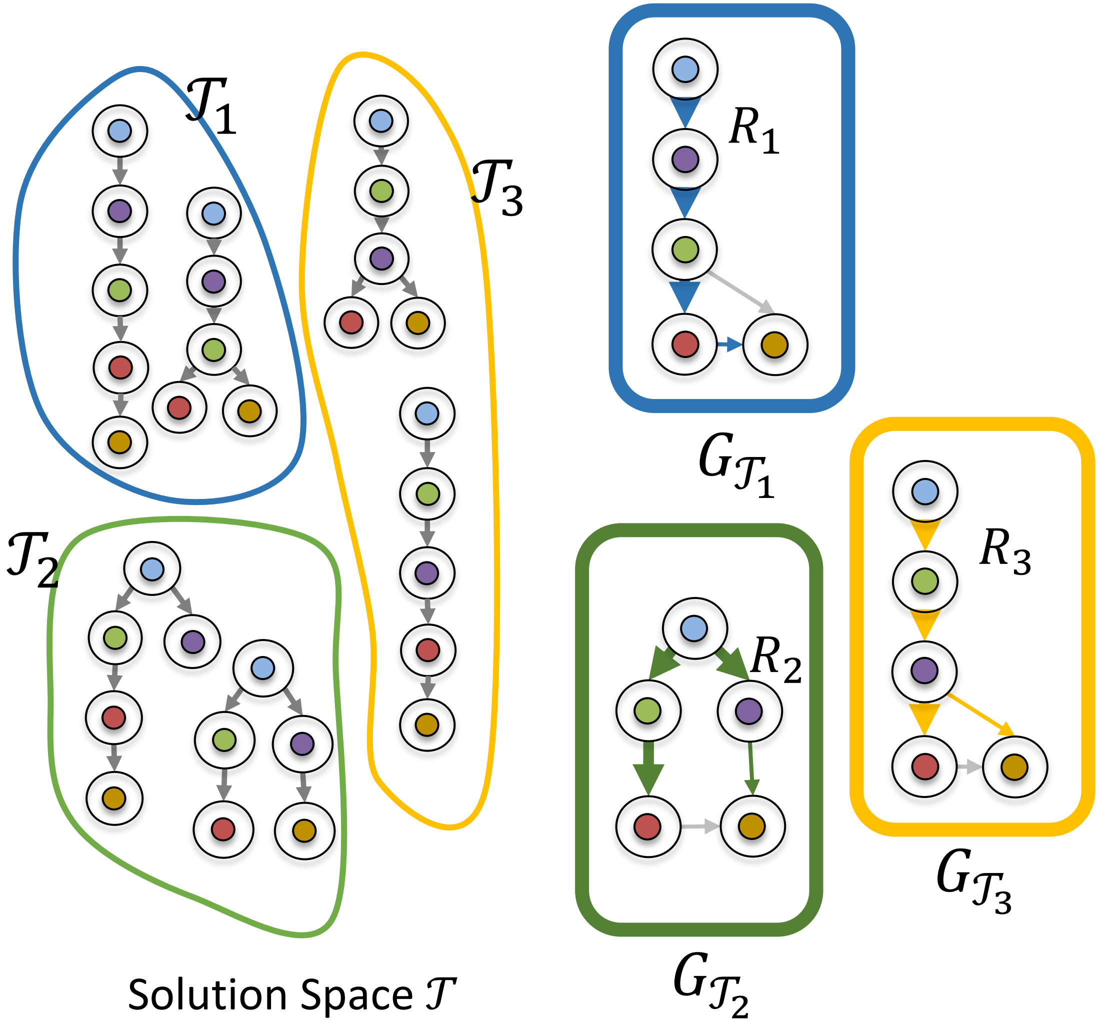
Combinatorial Characterization of Solutions to MCT

Single Consensus Trees (SCT): [Govek et al., ACM-BCB 2018]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$, find consensus tree R s.t.
$\sum_{i=1}^{n} d(T_i, R)$ is minimum

Theorem: [Govek et al., ACM-BCB 2018]
Max weight spanning arborescences
of parent-child graph $G_{\mathcal{T}}$ are solutions to SCT

Multiple Consensus Trees (MCT): [Aguse et al., ISMB 2019]
Given $\mathcal{T} = \{T_1, \ldots, T_n\}$ and $k > 0$, find surjective clustering
$\sigma : [n] \to [k]$ ~~and consensus trees~~ $\cancel{\mathcal{R} = \{R_1, \ldots, R_k\}}$
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum
where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$

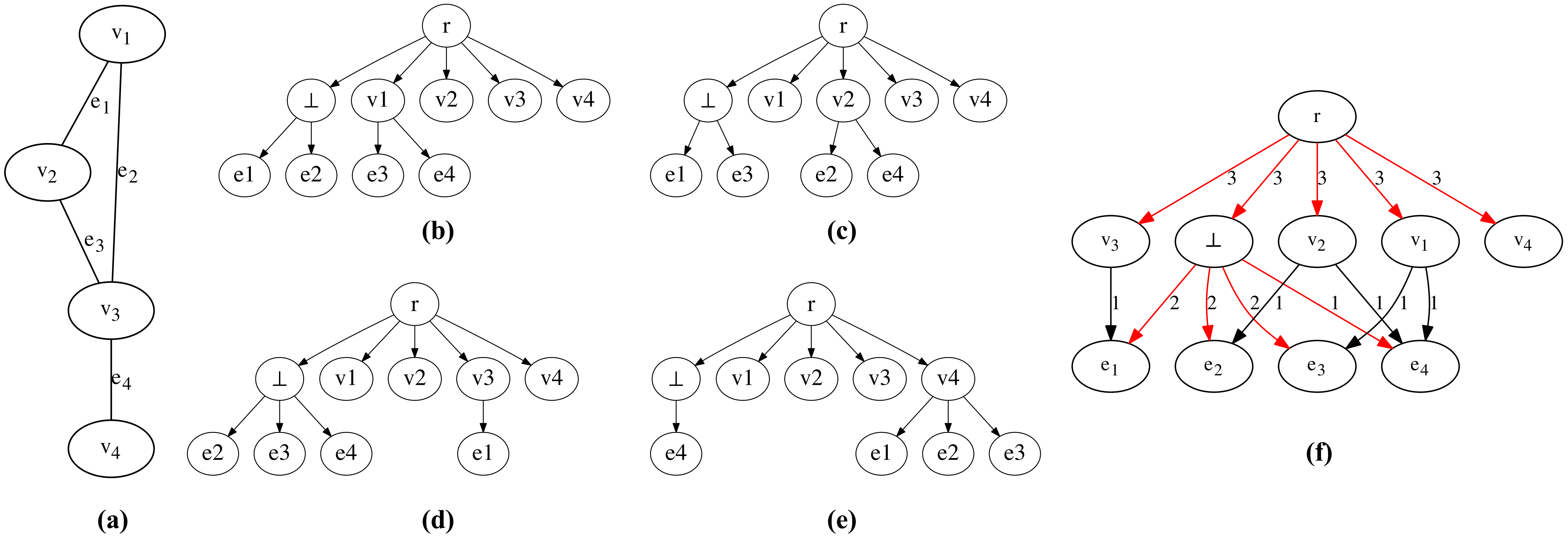
Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \to [k]$, MCT decomposes into
k independent SCT instances



Question: How to find σ^*?

Complexity

Multiple Consensus Trees (MCT):
Given $\mathcal{T} = \{T_1, \dots, T_n\}$ and $k > 0$, find surjective clustering $\sigma : [n] \to [k]$
s.t. $\sum_{i=1}^{n} d(T_i, R_{\sigma(i)})$ is minimum where $R_{\sigma(i)}$ is max weight spanning arborescence of $G_{\mathcal{T}_{\sigma(i)}}$



Theorem: MCT is NP-hard for general k (by reduction from CLIQUE).

Outline

- Problem Statement
    - Previous work
    - Problem statement
    - Combinatorial characterization of solutions
    - Complexity
- Method & Results
    - Exact algorithm
    - Heuristic algorithm
    - Model selection

Mixed Integer Linear Program

Theorem: MCT is NP-hard for general k (by reduction from CLIQUE).

$$
\begin{aligned}
\min\ & n(m-1) - \sum_{i=1}^{n}\sum_{s=1}^{k}\sum_{p=1}^{m}\sum_{q=1}^{m} w_{i,s,p,q} \\
\text{s.t.}\ & \sum_{s=1}^{k} x_{i,s} = 1 && \forall i \in [n] \\
& \sum_{i=1}^{n} x_{i,s} \geq 1 && \forall s \in [k] \\
& \sum_{p=1}^{m} z_{s,p} = 1 && \forall s \in [k] \\
& \sum_{q=1}^{m} y_{s,p,q} = 1 - z_{s,p} && \forall s \in [k], p \in [m] \\
& y_{s,p,q} \leq b_{p,q} && \forall s \in [k], p, q \in [m] \\
& \sum_{(p,q)\in\delta^-(U)} y_{s,p,q} + \sum_{p\in U} z_{s,p} \geq 1 && \forall s \in [k], U \subseteq [m] \\
& w_{i,s,p,q} \leq a_{i,p,q} && \forall i \in [n], s \in [k], p, q \in [m] \\
& w_{i,s,p,q} \leq x_{i,s} && \forall i \in [n], s \in [k], p, q \in [m] \\
& w_{i,s,p,q} \leq y_{s,p,q} && \forall i \in [n], s \in [k], p, q \in [m] \\
& w_{i,s,p,q} \geq 0 && \forall i \in [n], s \in [k], p, q \in [m] \\
& y_{s,p,q} \leq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} && \forall s \in [k], p, q \in [m] \\
& y_{s,p,q} \geq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} - \sum_{i=1}^{n} x_{i,s} + 1 && \forall s \in [k], p, q \in [m] \\
& \sum_{i=1}^{n} x_{i,s} \geq \sum_{i=1}^{n} x_{i,s+1} + 1 && \forall s \in [k-1] \\
& x_{i,s} \in \{0,1\} && \forall i \in [n], s \in [k] \\
& y_{s,p,q} \geq 0 && \forall s \in [k], p, q \in [m] \\
& z_{s,p} \geq 0 && \forall s \in [k], p \in [m]
\end{aligned}
$$

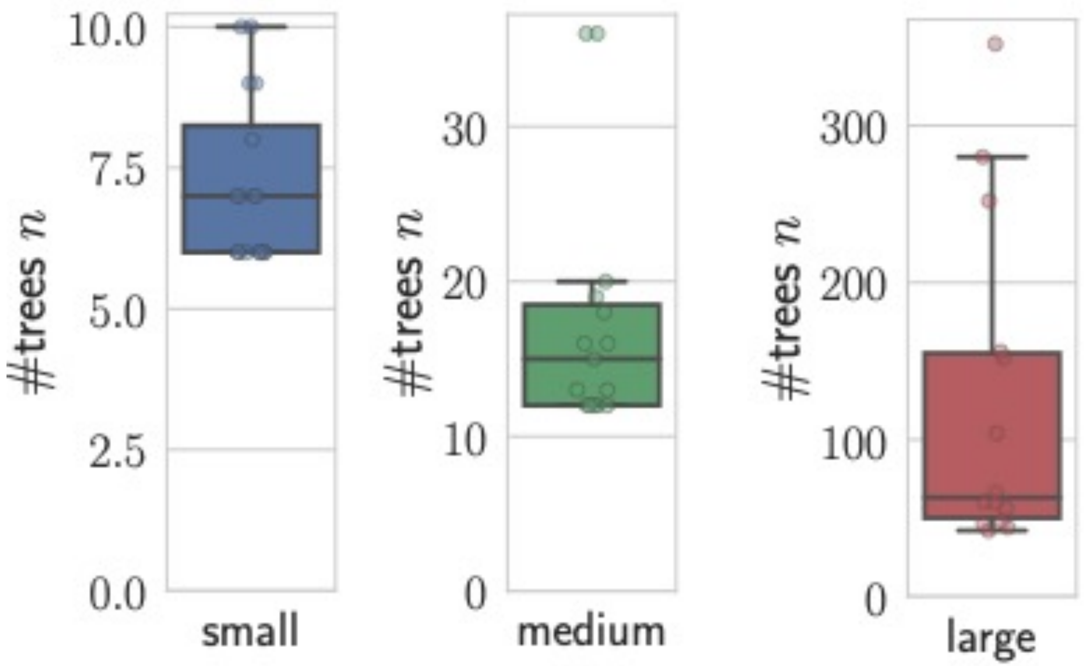
Mixed Integer Linear Program

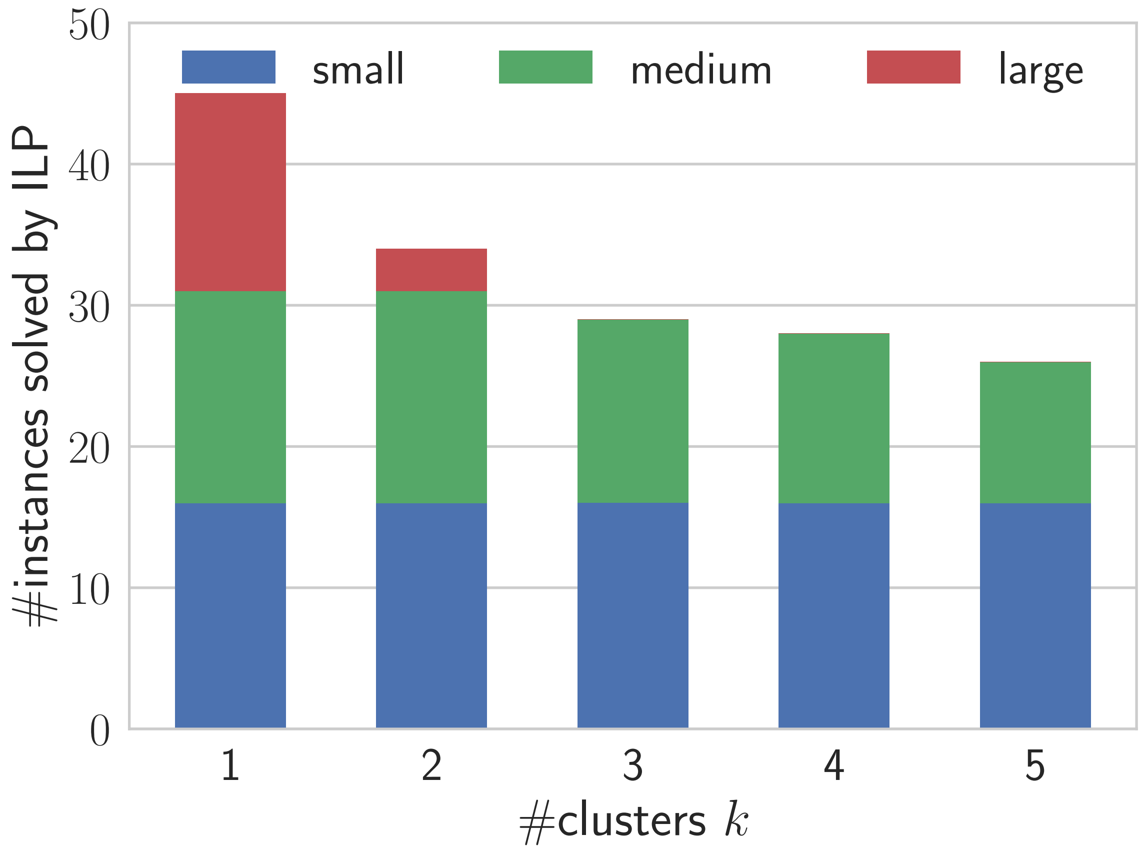
Theorem: MCT is NP-hard for general k (by reduction from CLIQUE).

$x_{i,s} \in \{0,1\}$	Tree T_i is assigned to cluster s
$y_{s,p,q} \geq 0$	Edge (p,q) is present in consensus tree R_s
$z_{s,p} \geq 0$	Vertex p is root of consensus tree R_s

$$
\begin{aligned}
\min\ & n(m-1) - \sum_{i=1}^{n}\sum_{s=1}^{k}\sum_{p=1}^{m}\sum_{q=1}^{m} w_{i,s,p,q} & \\
\text{s.t. } & \sum_{s=1}^{k} x_{i,s} = 1 & \forall i \in [n] \\
& \sum_{i=1}^{n} x_{i,s} \geq 1 & \forall s \in [k] \\
& \sum_{p=1}^{m} z_{s,p} = 1 & \forall s \in [k] \\
& \sum_{q=1}^{m} y_{s,p,q} = 1 - z_{s,p} & \forall s \in [k], p \in [m] \\
& y_{s,p,q} \leq b_{p,q} & \forall s \in [k], p,q \in [m] \\
& \sum_{(p,q)\in\delta^-(U)} y_{s,p,q} + \sum_{p\in U} z_{s,p} \geq 1 & \forall s \in [k], U \subseteq [m] \\
& w_{i,s,p,q} \leq a_{i,p,q} & \forall i \in [n], s \in [k], p,q \in [m] \\
& w_{i,s,p,q} \leq x_{i,s} & \forall i \in [n], s \in [k], p,q \in [m] \\
& w_{i,s,p,q} \leq y_{s,p,q} & \forall i \in [n], s \in [k], p,q \in [m] \\
& w_{i,s,p,q} \geq 0 & \forall i \in [n], s \in [k], p,q \in [m] \\
& y_{s,p,q} \leq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} & \forall s \in [k], p,q \in [m] \\
& y_{s,p,q} \geq \sum_{i=1}^{n} a_{i,p,q} x_{i,s} - \sum_{i=1}^{n} x_{i,s} + 1 & \forall s \in [k], p,q \in [m] \\
& \sum_{i=1}^{n} x_{i,s} \geq \sum_{i=1}^{n} x_{i,s+1} + 1 & \forall s \in [k-1] \\
& x_{i,s} \in \{0,1\} & \forall i \in [n], s \in [k] \\
& y_{s,p,q} \geq 0 & \forall s \in [k], p,q \in [m] \\
& z_{s,p} \geq 0 & \forall s \in [k], p \in [m]
\end{aligned}
$$

MILP does not scale well with k and n





Coordinate Ascent (akin to k-means)

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \to [k]$, MCT decomposes into k independent SCT instances

1. Fix clustering σ at random
2. Compute consensus tree R_s for each cluster s
3. Reassign each input trees T_i to cluster s where $d(T_i, R_s)$ is minimum
4. Go to 2

Coordinate Ascent (akin to k-means)

Proposition: [Aguse et al., ISMB 2019]
Given fixed clustering $\sigma : [n] \rightarrow [k]$, MCT decomposes into k independent SCT instances

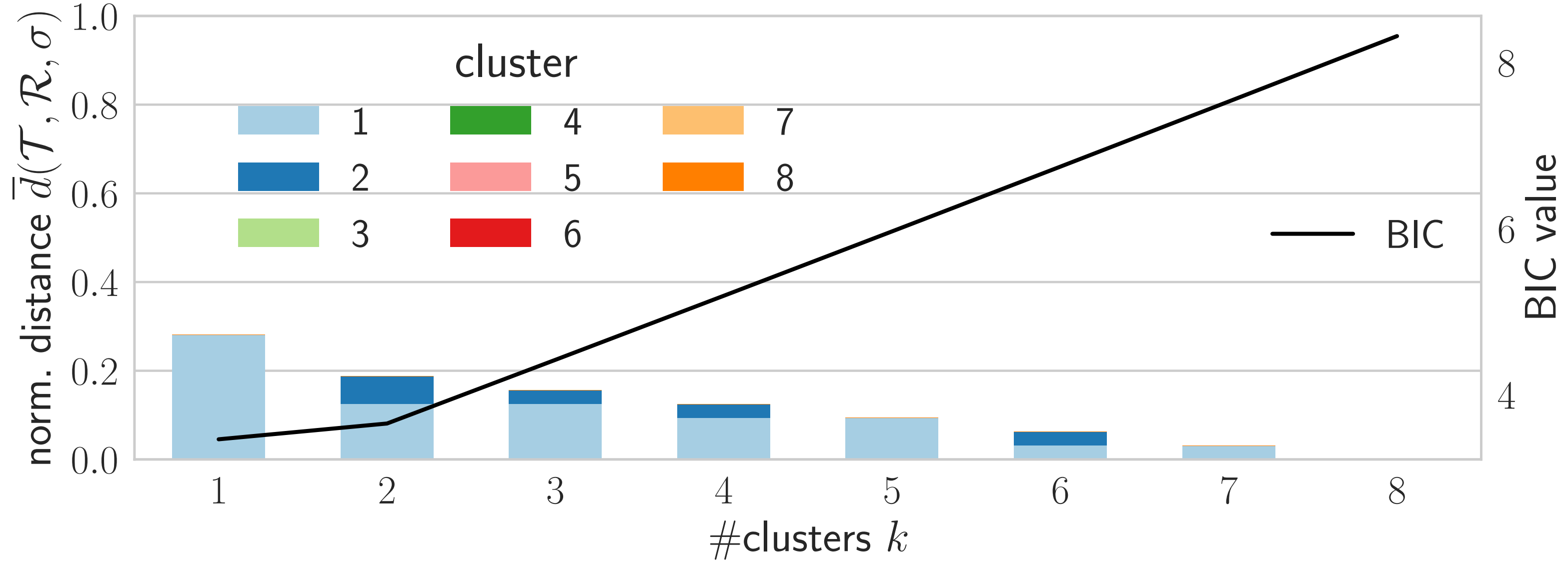
1. Fix clustering σ at random
2. Compute consensus tree R_s for each cluster s
3. Reassign each input trees T_i to cluster s where $d(T_i, R_s)$ is minimum
4. Go to 2

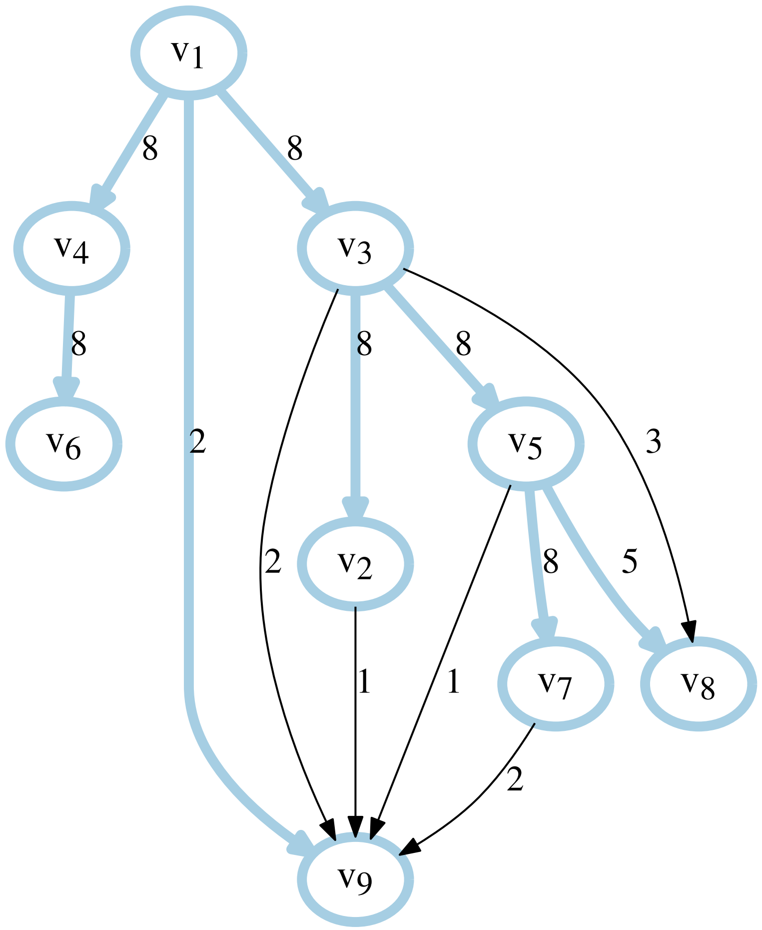
	#clusters k	MILP (1 h)	BF (1 h)	CA (1 h)	CA (100 r.)
small (16)	2	16	16	16	16
	3	16	16	16	16
	4	16	16	16	16
	5	16	14	16	16
medium (15)	2	15	13	15	15
	3	13	7	13	13
	4	12	0	12	12
	5	10	0	10	10
large (14)	2	3	0	3	3
	3	0	0	0	0
	4	0	0	0	0
	5	0	0	0	0

Bayesian Information Criterion

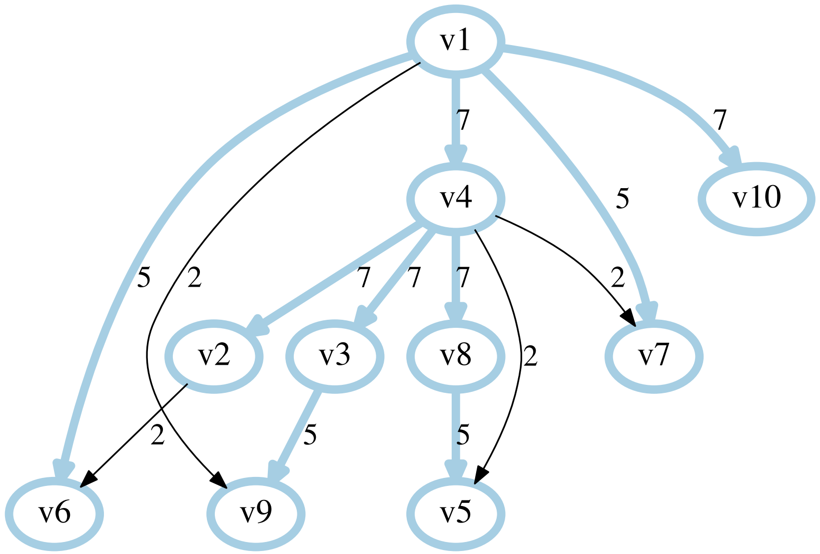
Jamal-Hanjani et al. (2017). *NEJM*.

Jamal-Hanjani et al. inferred 8 trees for patient CRUK0013



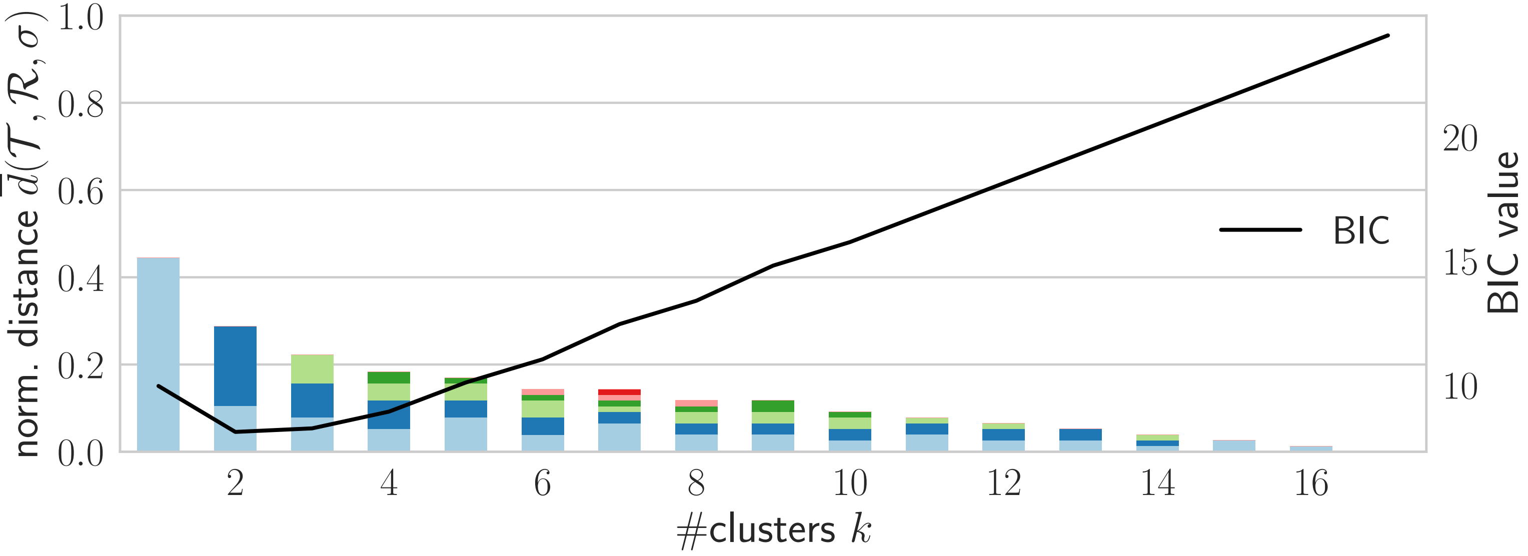


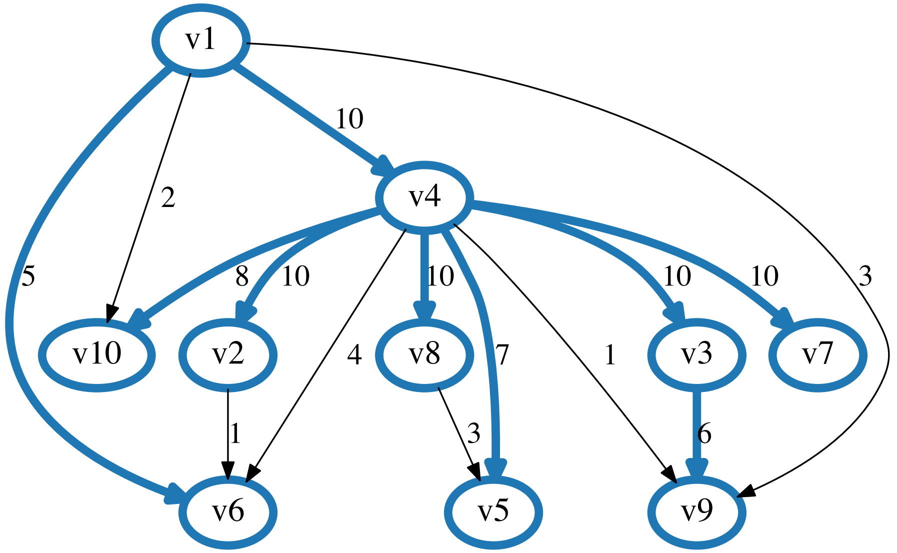
Bayesian Information Criterion



Jamal-Hanjani et al. (2017). *NEJM*.

Jamal-Hanjani et al. inferred 17 trees for patient CRUK0037





Conclusion

- Introduced the Multiple Consensus Tree (MCT) problem
- Characterized combinatorial structure of optimal solutions
- Showed that MCT is NP-hard
- Presented a mixed integer linear program
- Presented an efficient heuristic and showed that it finds optimal solutions
- Model selection for the number of clusters

Future directions

- Relax infinite sites assumption
- Use medoids rather than centroids