Crowdsourcing Taxonomies

Dimitris Karampinas & Peter Triantafillou

Computer Engineering and Informatics Department - University of Patras

INTRODUCTION

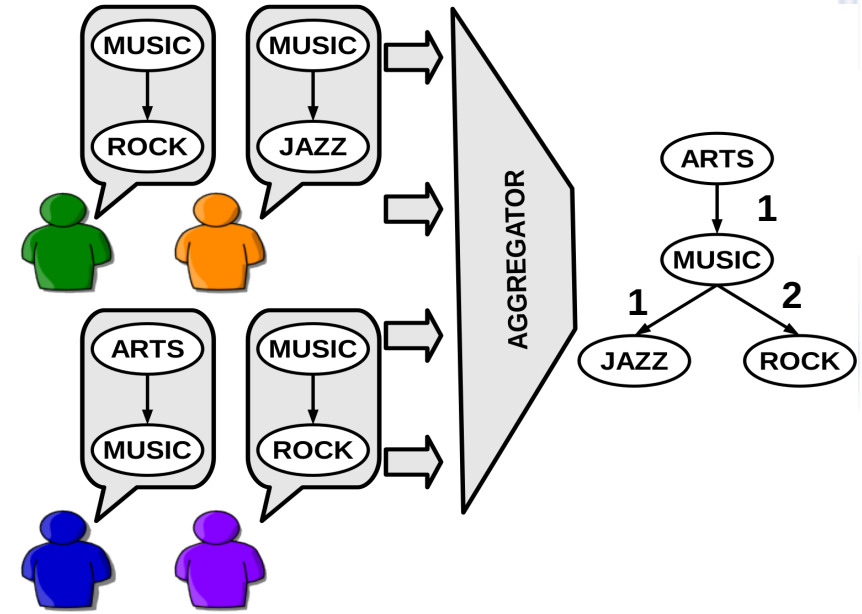
- Social Media & Tagging
  - Users organize their own data
  - Users fulfill the need of communication
- Crowdsourcing (Crowd + Outsourcing)
  - Massive collaboration
  - Collective Intelligence
- Taxonomies
  - Index and search web content
  - Need for experts (maintenance overhead!)

ROADMAP

- Introduction
- Model & Theory
- Algorithm
- Experimentation
- Conclusions

MODEL

- Extended Tagging: 'Vote' ($tag_a \rightarrow tag_d$)
- Vote Cardinality – Weights
- Challenges & Goals
  - Individuals are prone to errors
  - Incomplete (structural) information
  - Incremental process / Maintainability
- Vote Types
  - Ancestor → Descendant
  - Descendant → Ancestor
  - Crosslinks



THEORY

- Invariants
  - Integrity of Tree Properties
  - Maximum Votes Satisfiability
- Trivial Maximum Spanning Tree algorithms do not suffice
- Problem Definition:

$$INPUT: Complete\ graph\ G=(V,E), weight\ w(e) \in \mathbb{Z}^0\ for\ each\ e \in E$$

$$OUTPUT: A\ spanning\ tree\ T\ for\ G\ such\ that, if\ W(u,v)\ denotes\ the\ \text{sum}$$

$$of\ weights\ of\ the\ edges\ on\ the\ path\ joining\ u\ \text{and}\ v\ \text{in}\ T, then\ find\ B\ where:$$

$$B = max\left(\sum_{u,v \in V} W(u,v)\right)$$

(B represents the maximum number of votes that is satisfied)

THEORY

- Invariants
  - Integrity of Tree Properties
  - Maximum Votes Satisfiability
- Trivial Maximum Spanning Tree algorithms do not suffice
- Problem Definition:

$$INPUT: Complete\ graph\ G=(V,E), weight\ w(e) \in \mathbb{Z}^0\ for\ each\ e \in E$$

$$OUTPUT: A\ spanning\ tree\ T\ for\ G\ such\ that, if\ W(u,v)\ denotes\ the\ \text{sum}\ of\ weights\ of\ the\ edges\ on\ the\ path\ joining\ u\ \text{and}\ v\ \text{in}\ T, then\ find\ B\ where:$$

$$B = max\left(\sum_{u,v \in V} W(u,v)\right)$$

(B represents the maximum number of votes that is satisfied)

NP-Hard !
Optimum Communication Spanning Tree (Garey & Johnson)

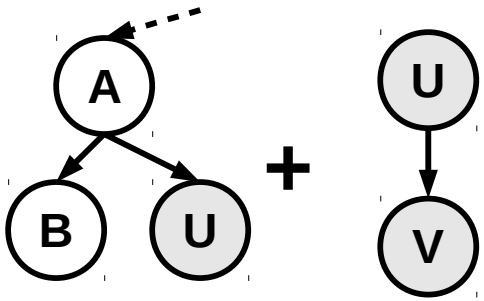
ROADMAP

- Introduction
- Model & Theory
- Algorithm
- Experimentation
- Conclusions

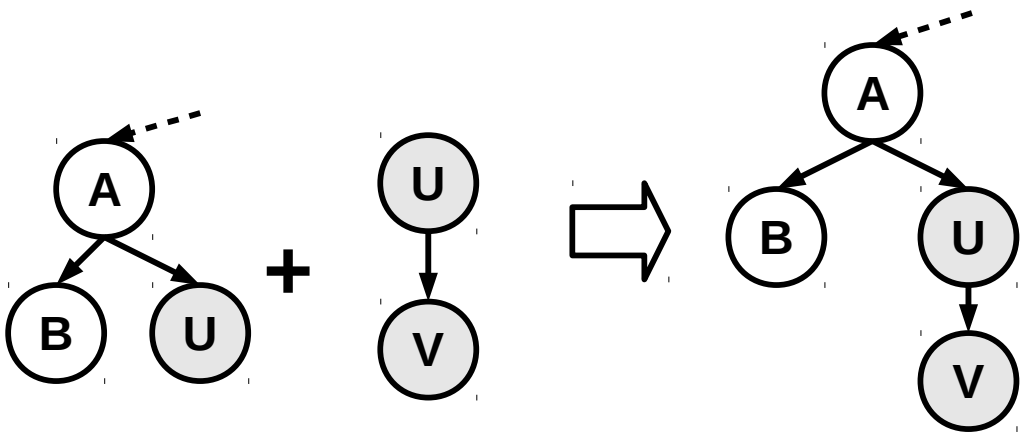
Vote Classification Routine

```
For all incoming votes (tag_anc→tag_des)

   Node u ←  hash(tag_anc)
   Node v ←  hash(tag_des)

   if (only u exists)
      attach_new_child(u, v);

   else if (none exists)
      create_new_tree(u, v);

   else if (only v exists)
      merge(u, v);

   else (both u AND v exist)
      LeastCommonAncestor(u, v);
         expand_vertically(u, v);
         backedge_resolution(u, v);
```

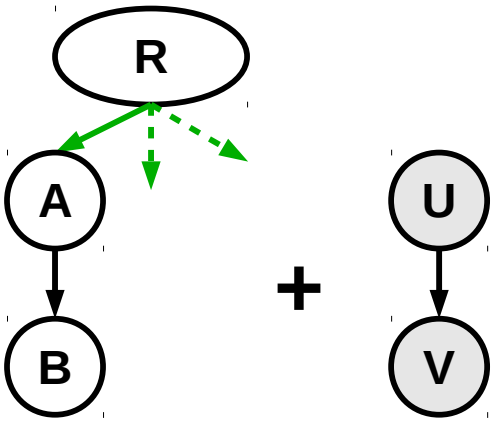
TRSFRM: ATTACH NEW CHILD



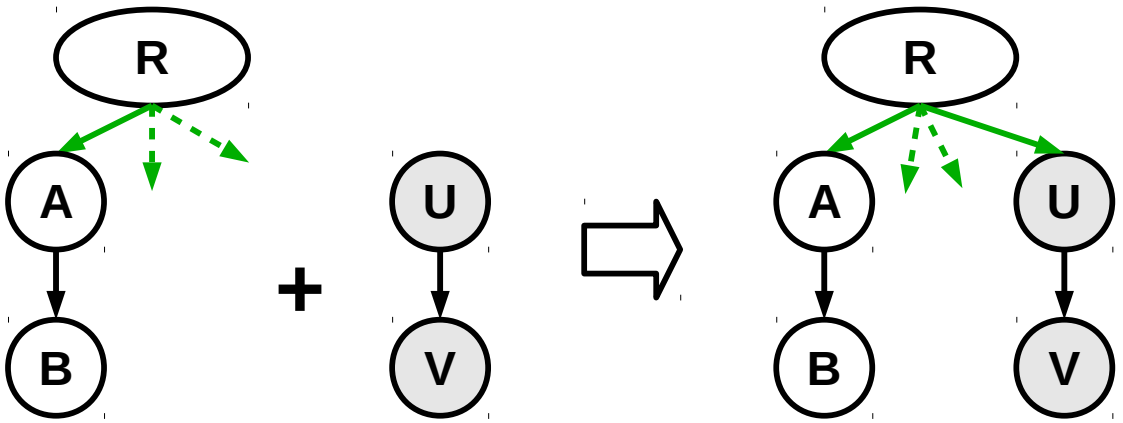
TRSFRM: ATTACH NEW CHILD



TRSFRM: CREATE NEW TREE

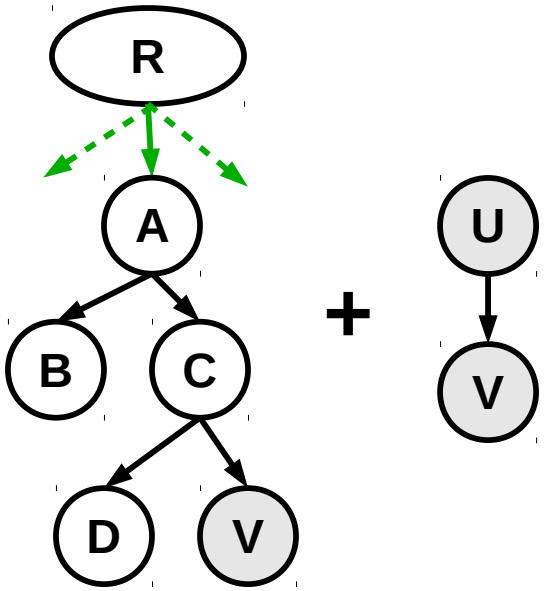


TRSFRM: CREATE NEW TREE

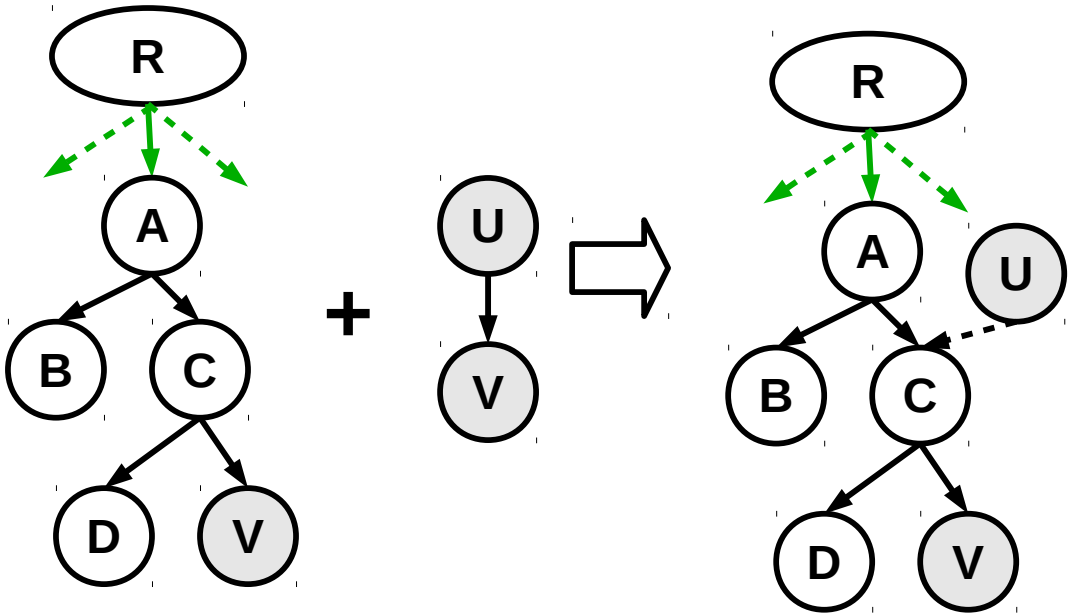


Shortcut: An artificial Parent→Child link that is not an explicit user generated relation, so its weight equals 0 and it is utilized in order to preserve structural continuation.

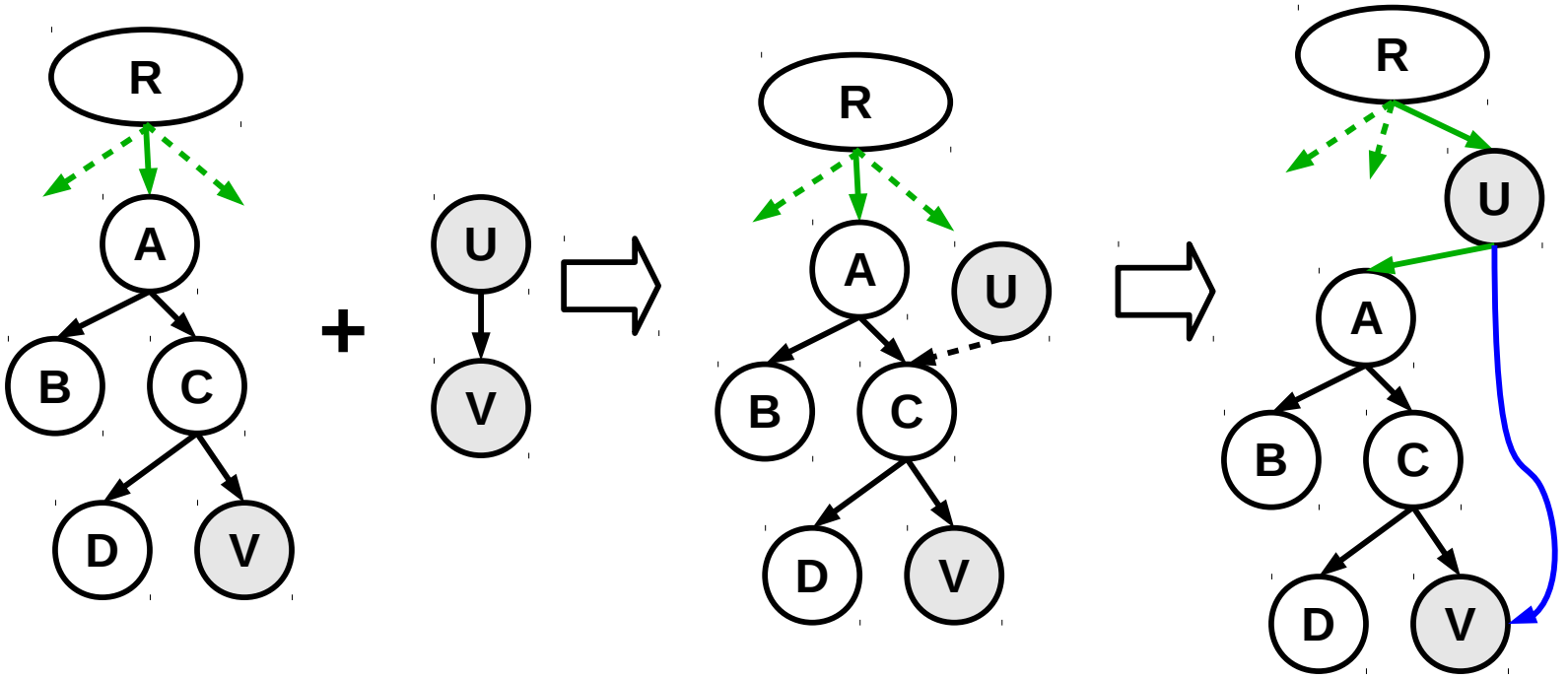
TRSFRM: MERGE



TRSFRM: MERGE

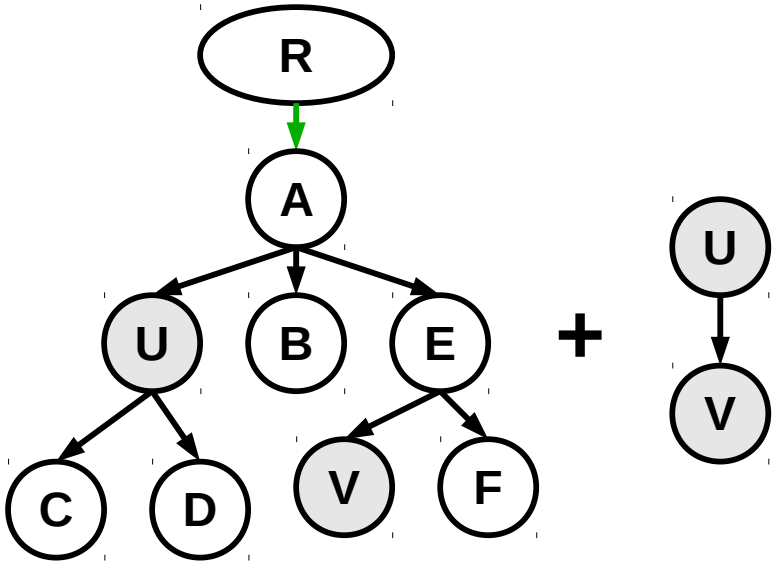


TRSFRM: MERGE

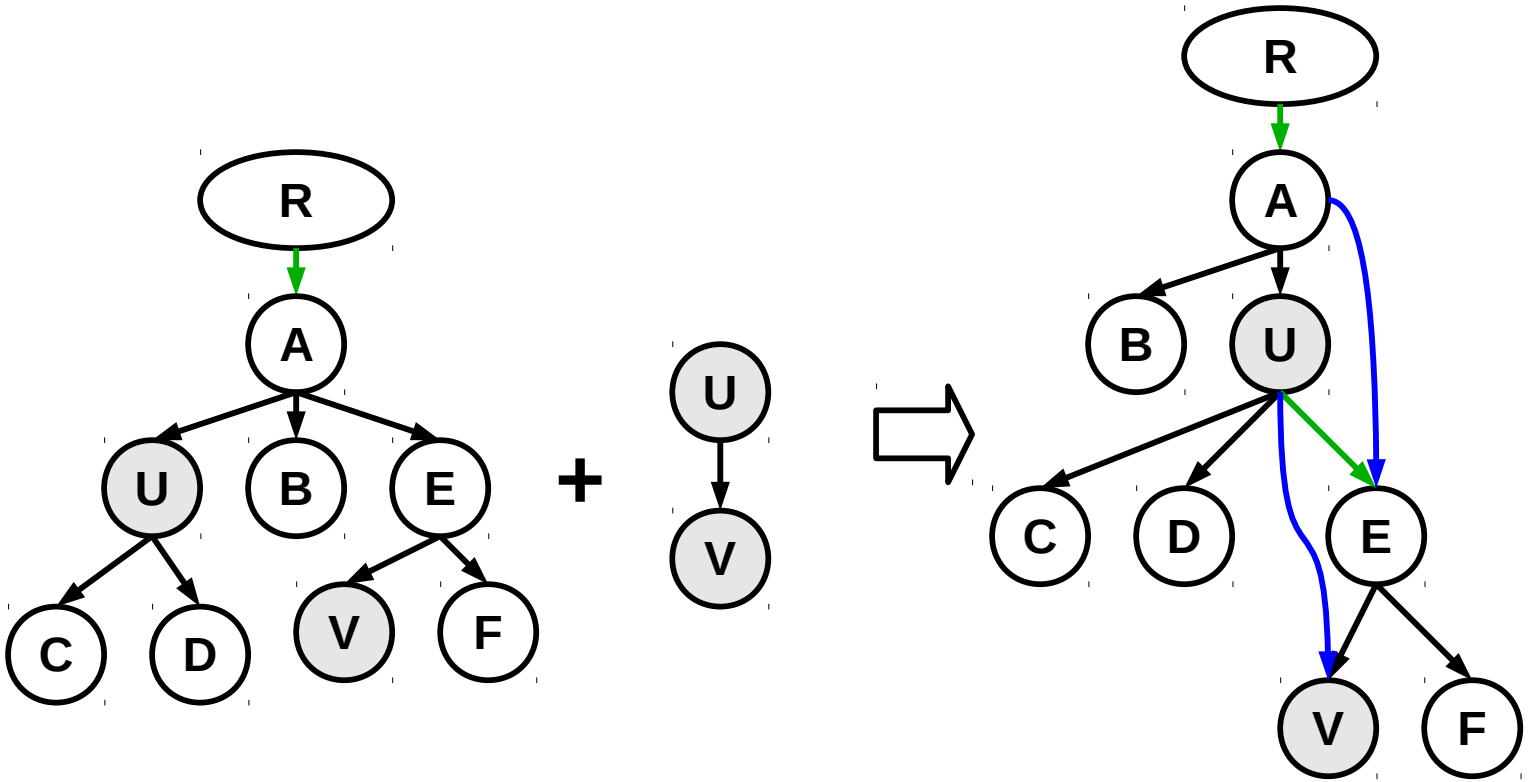


Forward Edge: A latent relation between two nodes. The source node is an ancestor in the taxonomy and the target is a descendant. Forward edges do not refer to Parent → Child links and remain hidden since they violate tree's properties.

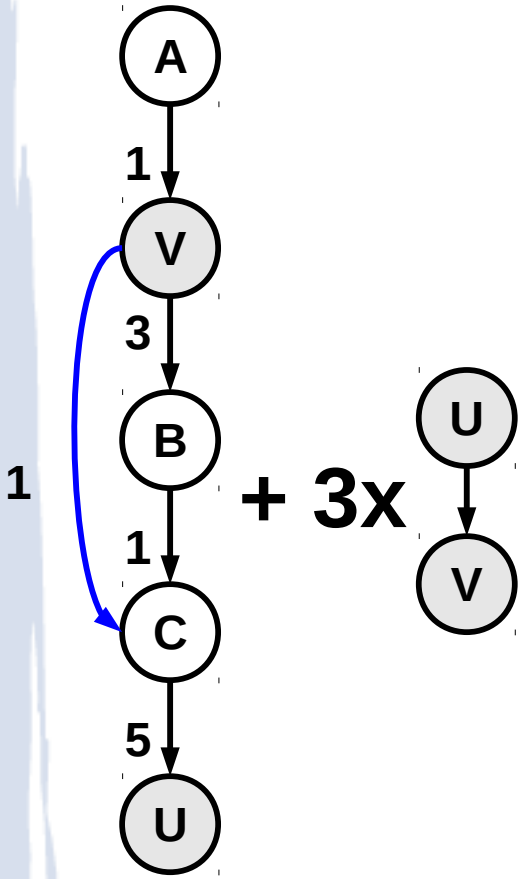
TRSFRM: EXPAND VERTICALLY



TRSFRM: EXPAND VERTICALLY

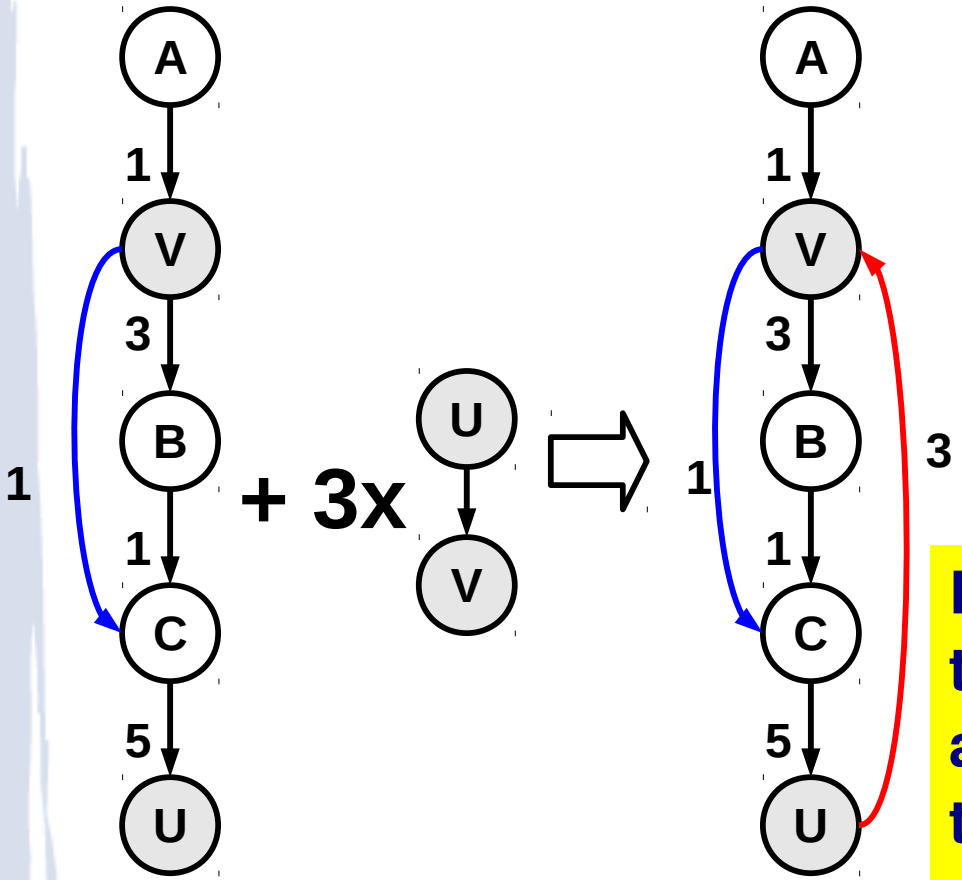


TRSFRM: BACKEDGE RESOLUTION



Initial State S

TRSFRM: BACKEDGE RESOLUTION

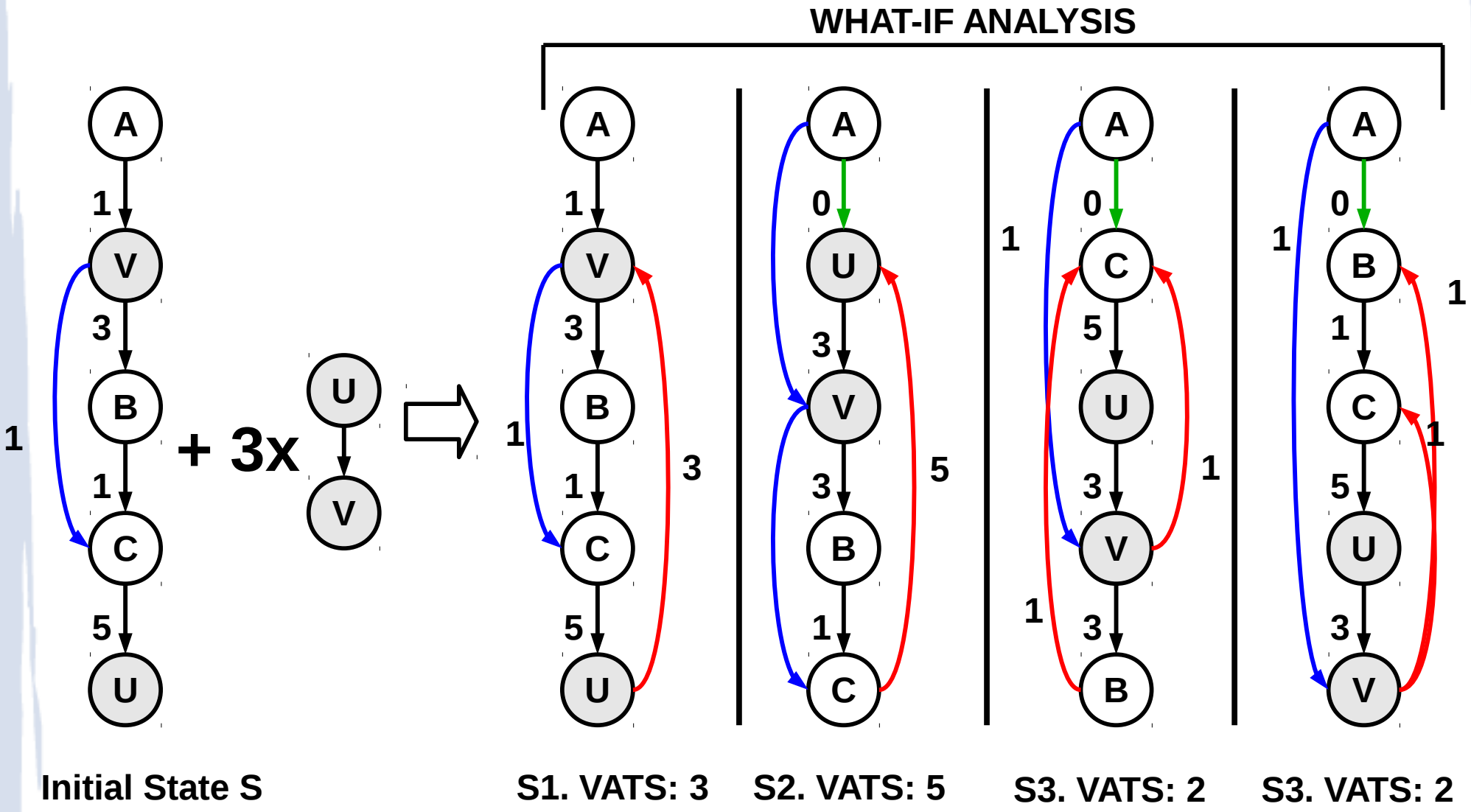


Initial State S

Backedge: A latent relation between two nodes. The source node appears as a descendant in the current taxonomy whereas the target as an ancestor. Backedges remain hidden since they violate the tree's properties.

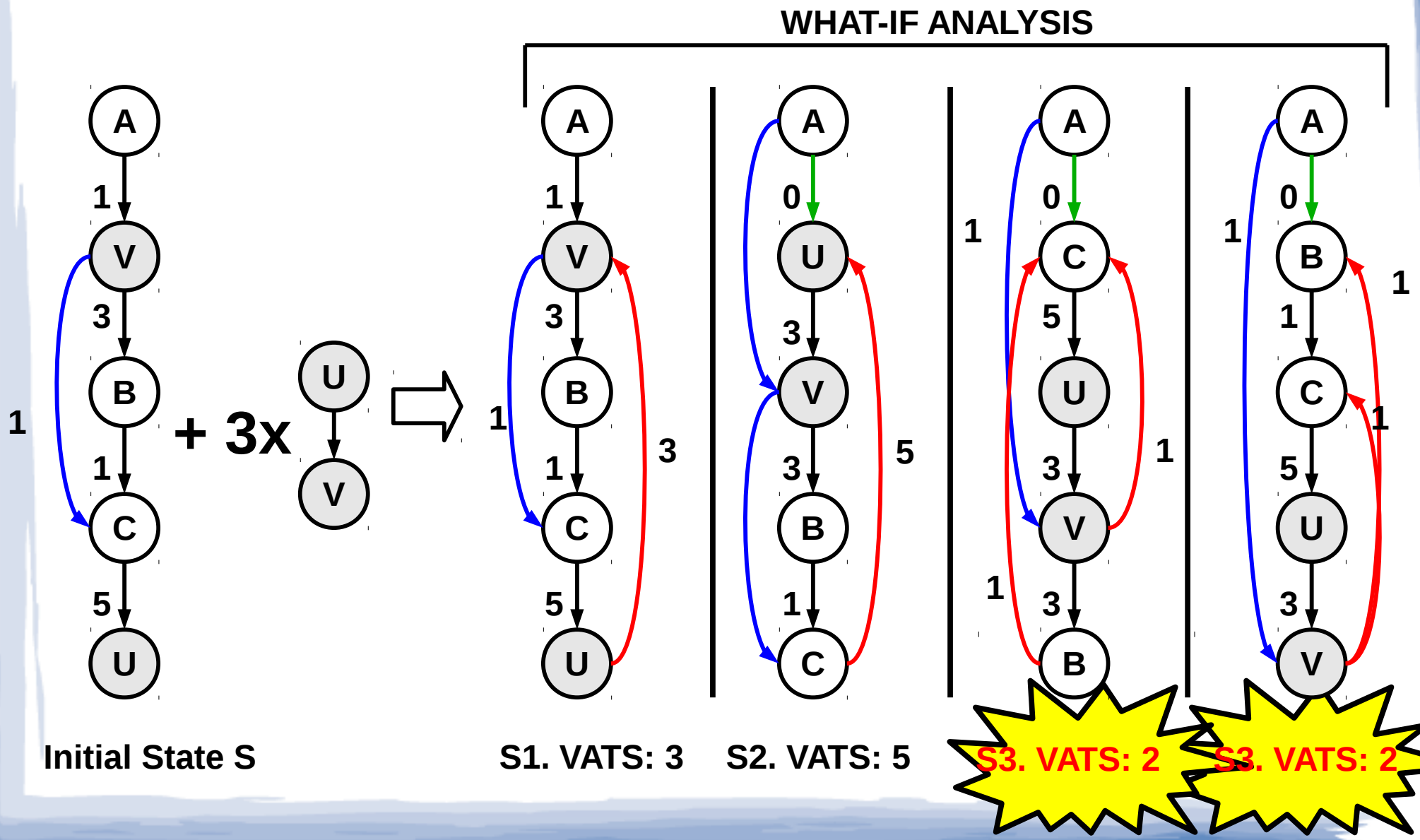
TRSFRM: BACKEDGE RESOLUTION

VATS: Votes Against This State



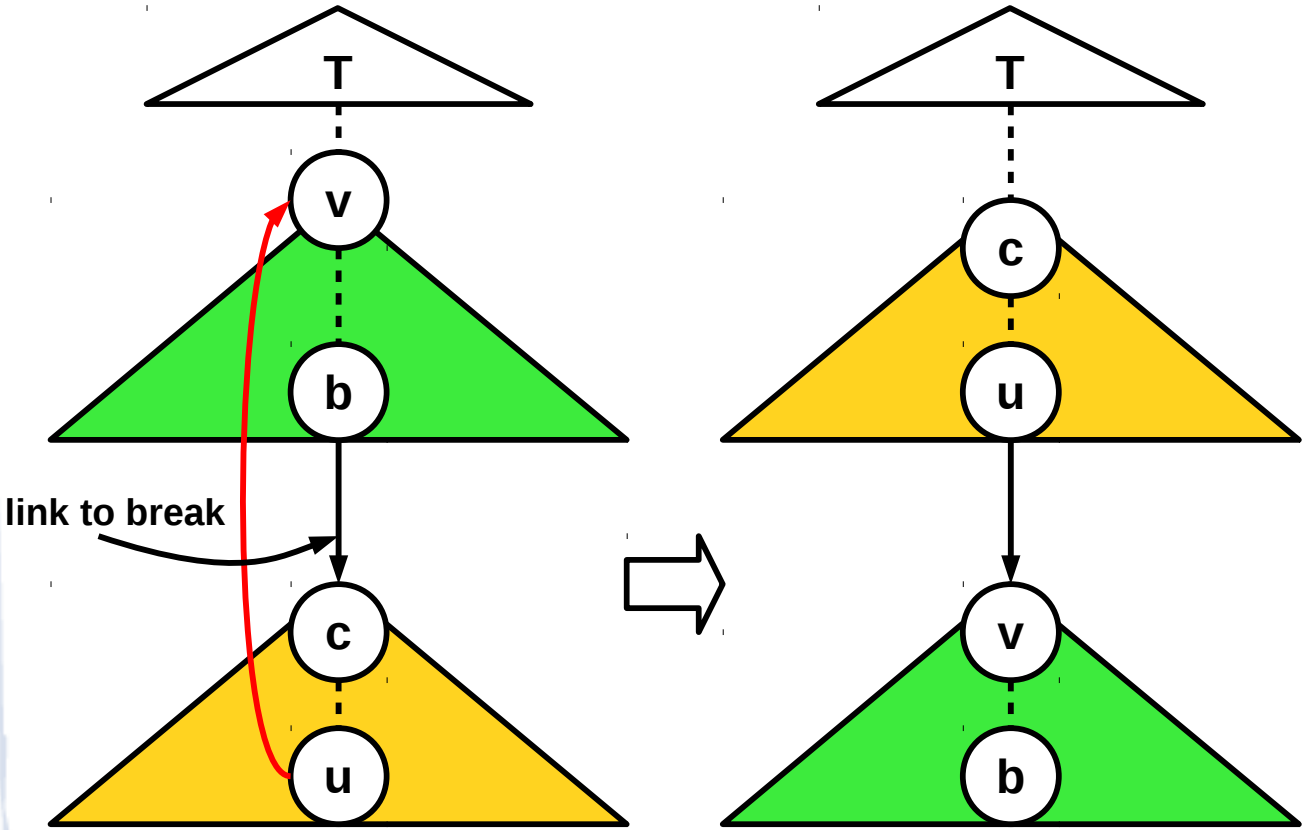
TRSFRM: BACKEDGE RESOLUTION

VATS: Votes Against This State



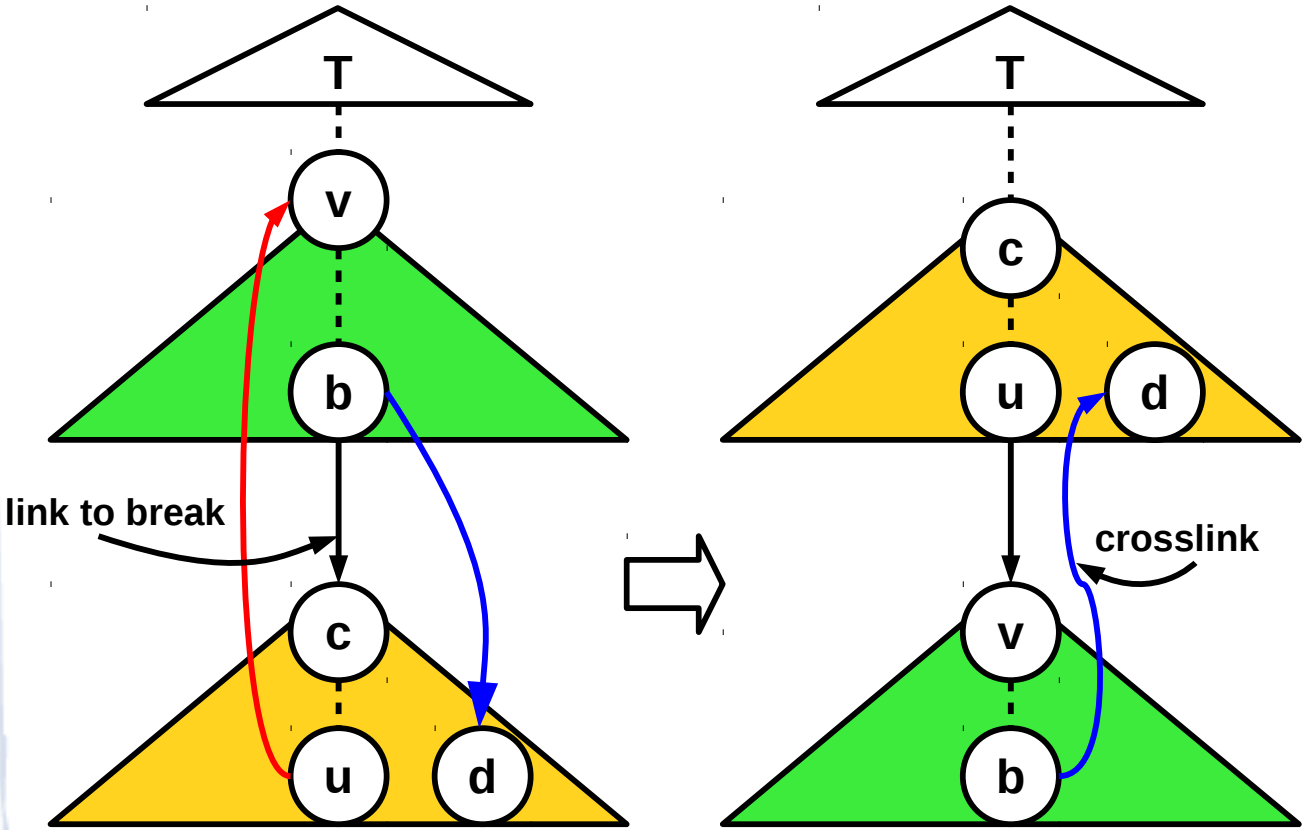
TRSFRM: BACKEDGE RESOLUTION

The Big Picture



TRSFRM: BACKEDGE RESOLUTION

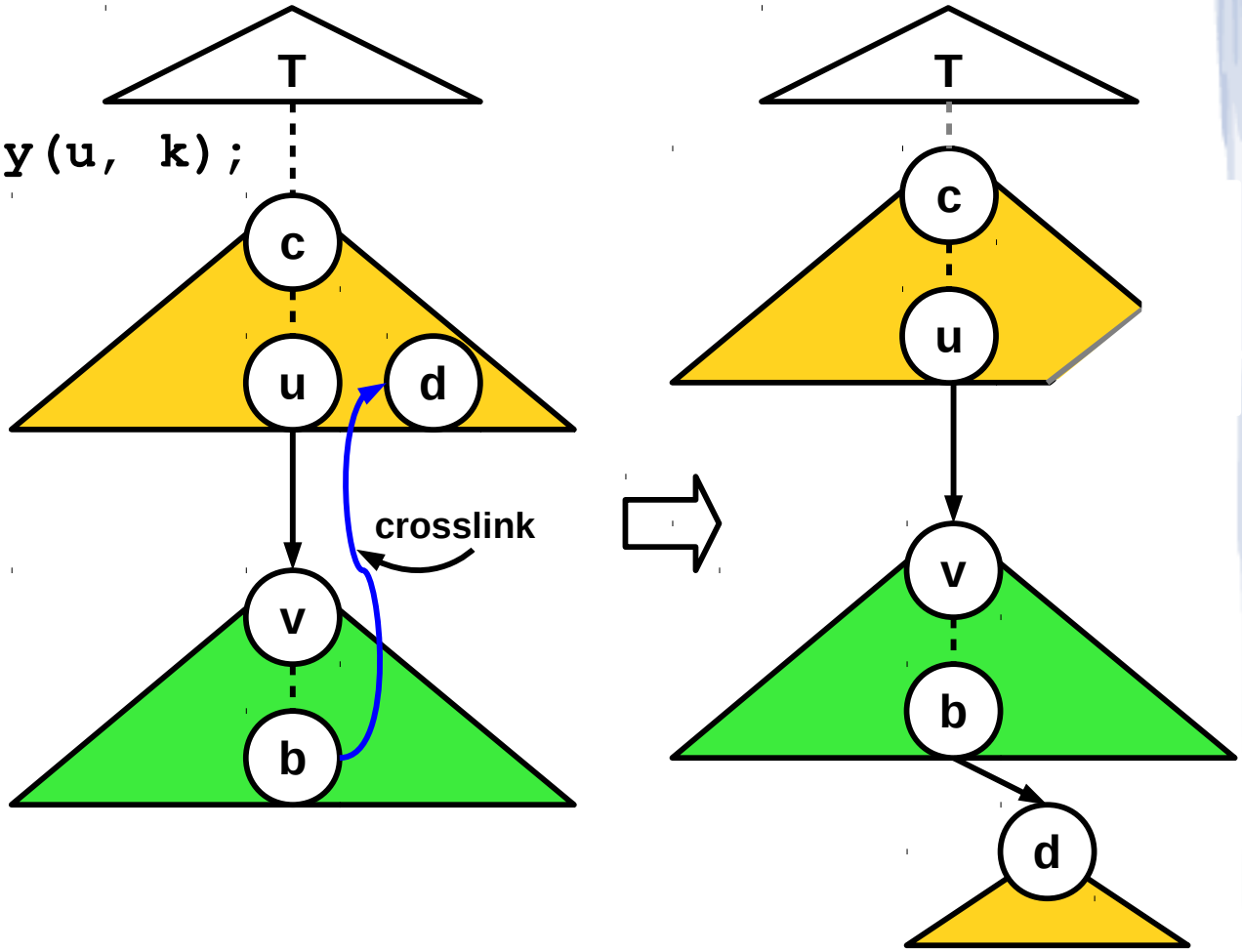
Follow ups



TRSFRM: BACKEDGE RESOLUTION

Follow ups

```
for all nodes u ∈ {v...b}
    for all u->fwd_edge(k)
        if (u == LCA(u, k))
            expand_vertically(u, k);
```



ROADMAP

- Introduction
- Model & Theory
- Algorithm
- Experimentation
- Conclusions

DATASET & METRICS

Dataset: ACM Computing Classification System

- 1473 Nodes
- 3824 (Ancestor → Descendant) Relations

Metrics:

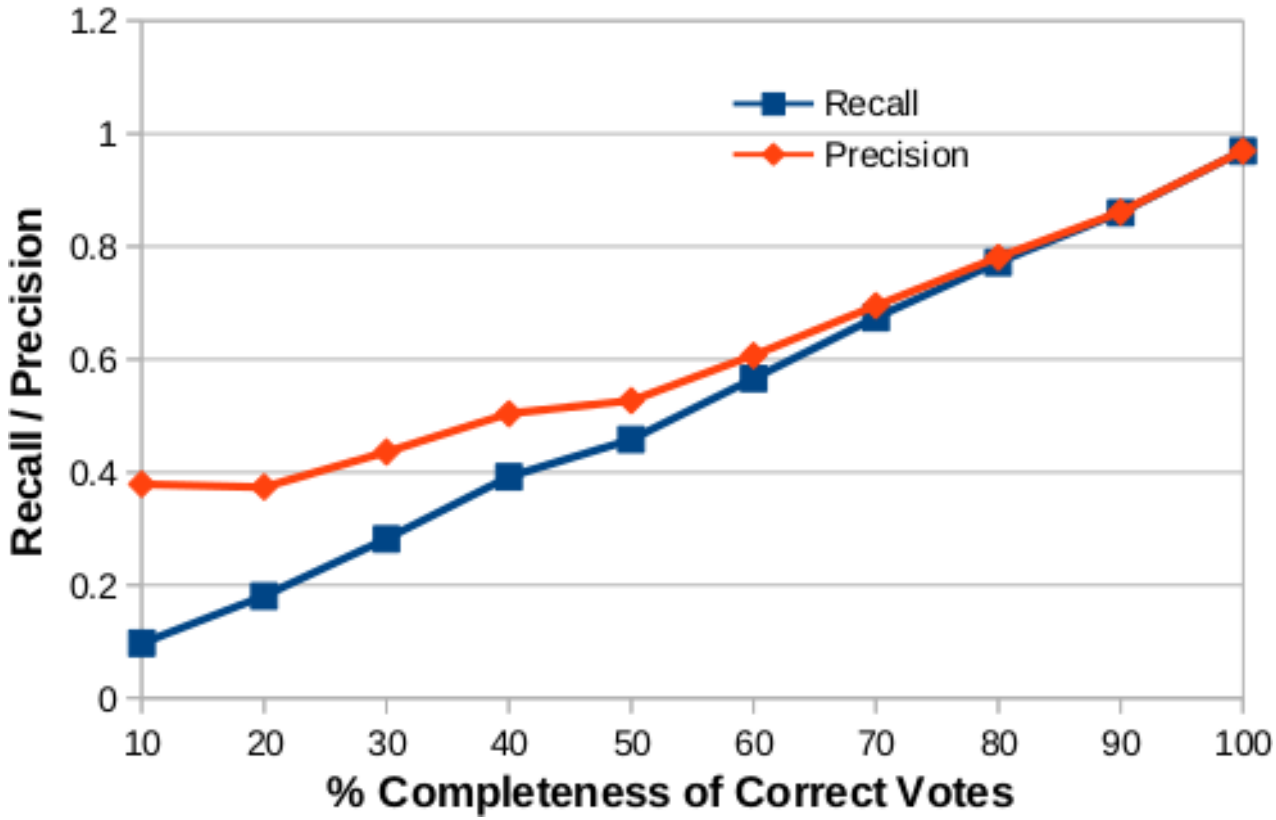
- To → Original Tree
- T → Tree derived from Crowd
- Ro → Set of Parent → Child relations in To
- R → Set of Parent → Child relations in T

$$Recall = \frac{|R \cap Ro|}{|Ro|} \qquad Precision = \frac{|R \cap Ro|}{|R|}$$

$$FScore = \frac{2\,x\,Recall\,x\,Precision}{Recall + Precision}$$

SYNTHETIC EXPERIMENT - 1

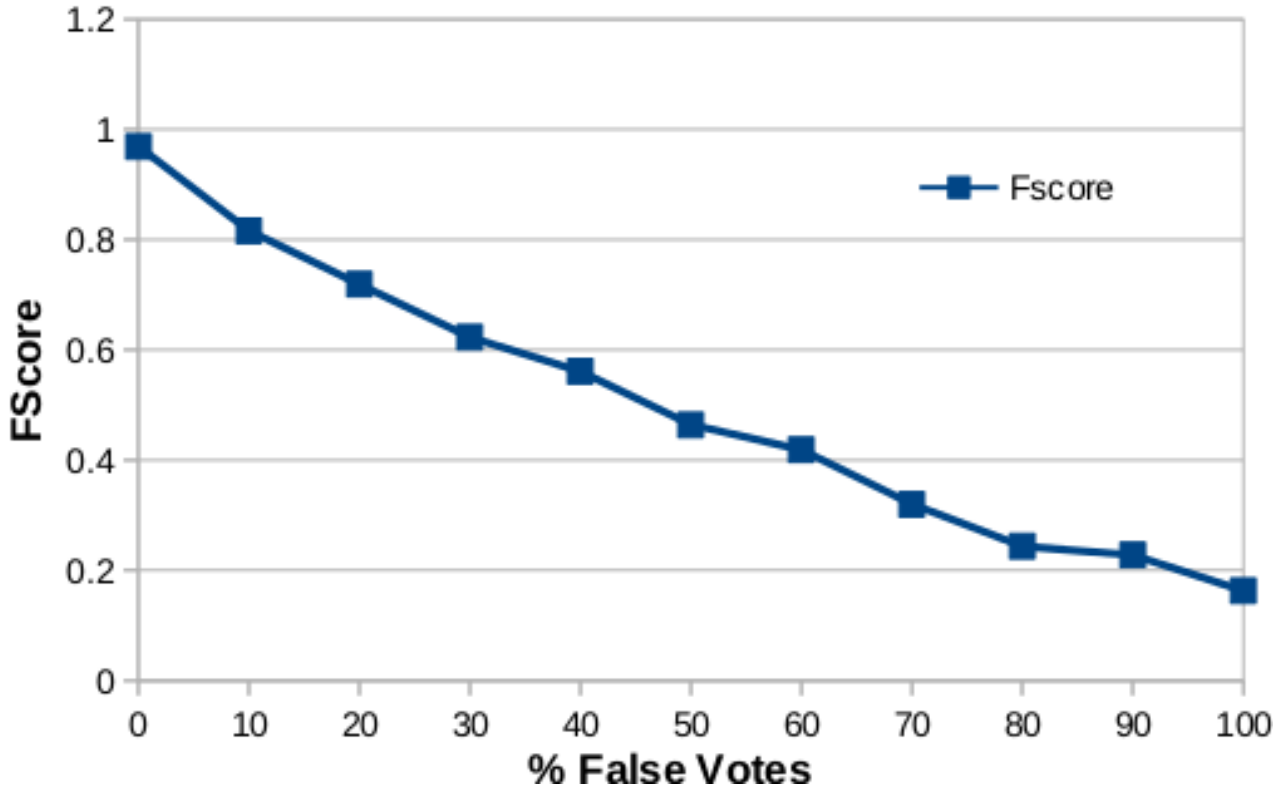
Setup: On every iteration reduce by 10% the number of Correct Votes



SYNTHETIC EXPERIMENT - 2

Setup: 100% Correct Votes

On every iteration increase by 10% the number of False Votes



CROWDSOURCING EXPERIMENT

- 245 Nodes
- 620 Ancestor → Descendant Relations (+ inverted counterparts)
- Preprocessing: Remove misleading concepts ('Misc', 'General')
- HIT: 25 pairs of relations presented in groups of 5
- 3-day Period

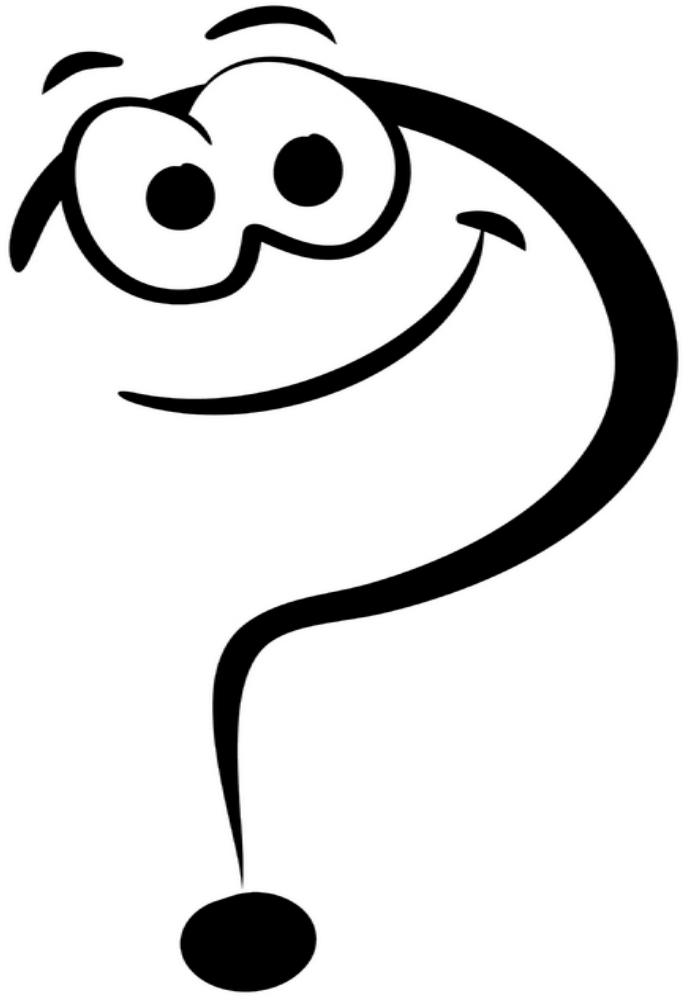
Users	102
Total Votes	2155
Correct Votes	1501
False Votes	435
Undefined	229
Rel. without Votes	24
Rel. without Correct Votes	33
Rel. without False Votes	302
Max Votes on a Rel.	10
FScore	0.486
Predicted FScore	0.519

ROADMAP

- Introduction
- Model & Theory
- Algorithm
- Experimentation
- Conclusions

CONCLUSIONS

- Novel idea for a Taxonomy development in Crowdsourcing environments
- Challenges & Model
- Problem definition (NP-Hardness)
- Algorithm
  - Online
  - Copes with lack of input & user conflicts
- Synthetic & Real world crowdsourcing experiments



Thank you crowd...