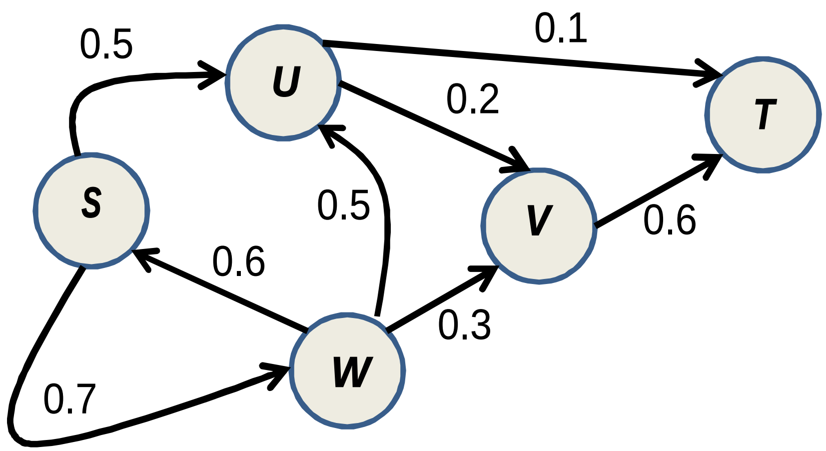
# Fast Reliability Search in Uncertain Graphs

**Arijit Khan, Francesco Bonchi, Aristides Gionis, Francesco Gullo**

Systems Group, ETH Zurich
Yahoo Labs, Spain
Aalto University, Finland

# Uncertain Graphs




Uncertain Graph

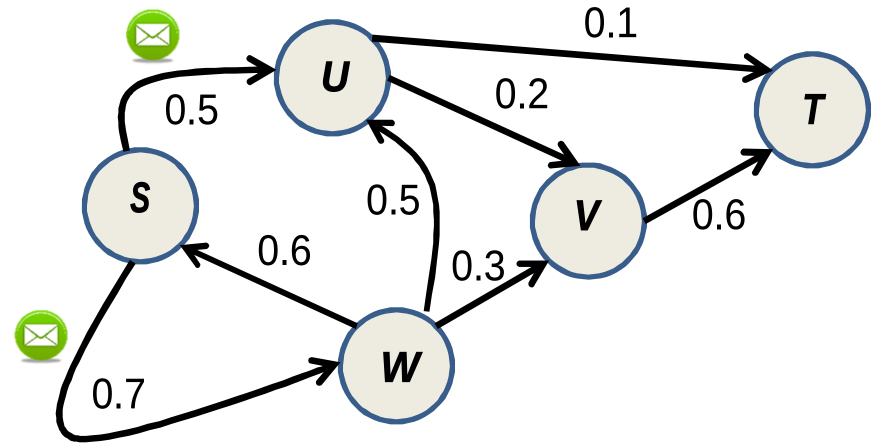
- Social Network
- Traffic Network
- Ad-hoc Mobile Network
- Protein-interaction Network

# Motivation

- **Mobile Ad-hoc Network:** find the set of sink nodes where a source node can deliver a packet with high probability

- **Traffic Network:** find a set of target locations reachable from a source location with high probability

- **Social Network:** find a set of users who could be influenced with high probability by a target user



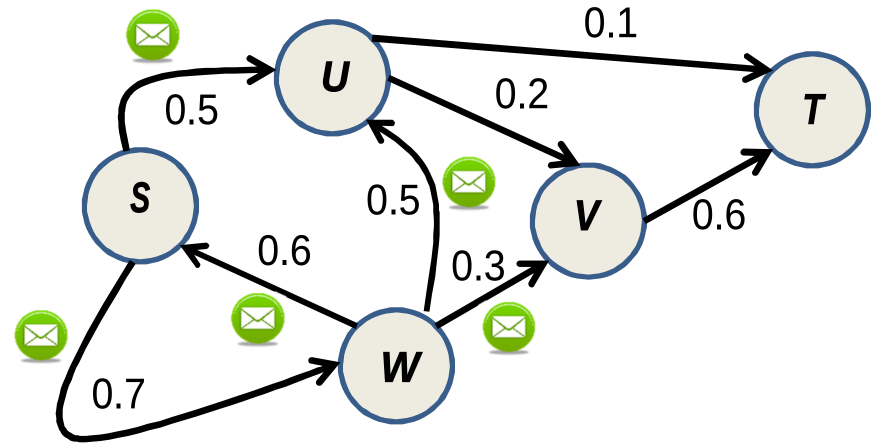

Packet Delivery Probability in Mobile Ad-hoc Network

# Motivation

- **Mobile Ad-hoc Network:** find the set of sink nodes where a source node can deliver a packet with high probability

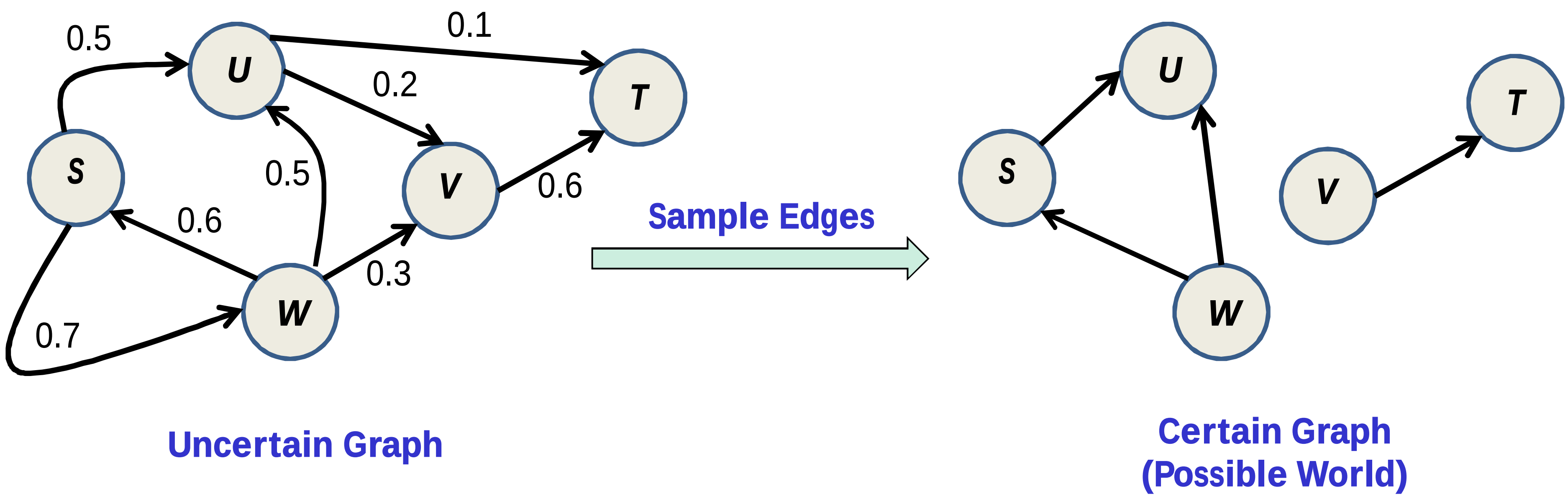
- **Traffic Network:** find a set of target locations reachable from a source location with high probability

- **Social Network:** find a set of users who could be influenced with high probability by a target user

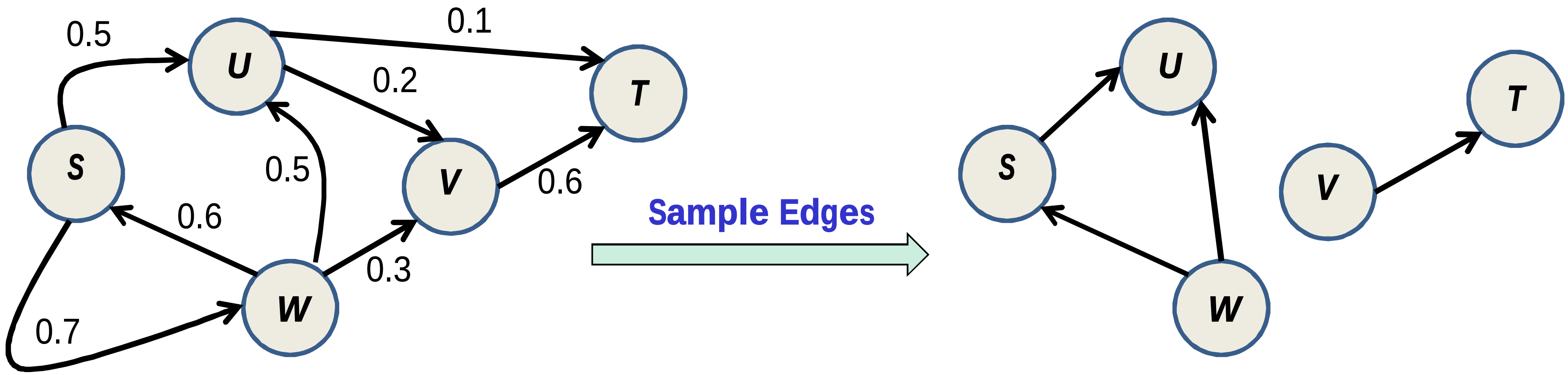



**Packet Delivery Probability in Mobile Ad-hoc Network**

# Reliability in Uncertain Graphs



# Reliability in Uncertain Graphs




Uncertain Graph

Certain Graph
(Possible World)

$$\Pr(G) = \prod_{a \in A_G} p(a) \prod_{a \in A \setminus A_G} (1 - p(a))$$

$$R(S,t) = \sum_{G \sqsubseteq \mathcal{G}} \underline{P_G(S,t)} \Pr(G)$$

Identity Function

# Reliability Search in Uncertain Graphs

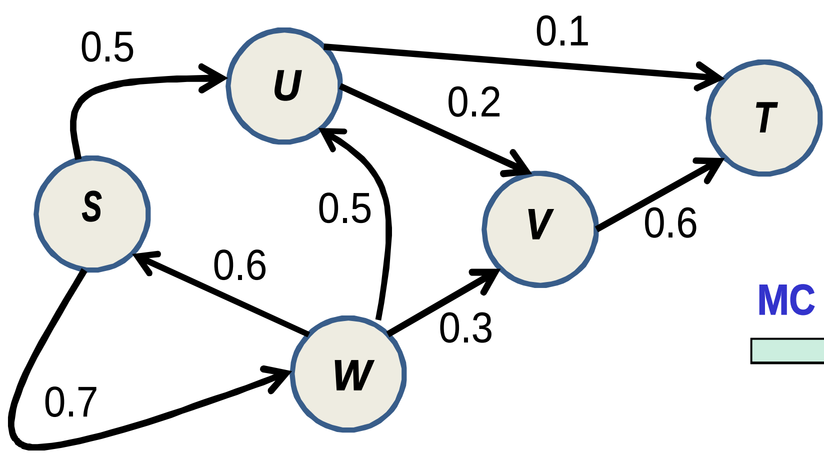
Given an uncertain graph G, a probability threshold $\eta \in (0, 1)$, and a source node S in G, find all nodes in G that are reachable from S with probability greater than or equal to threshold $\eta$

- #P - complete

# Related Work

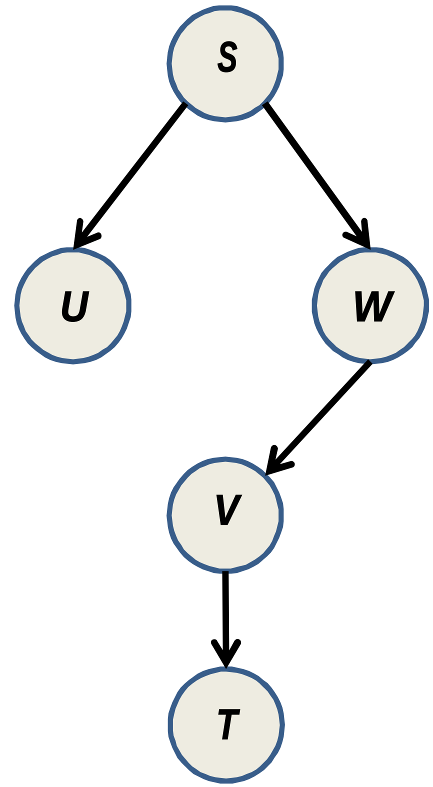
- Two-terminal reliability
- All-terminal reliability
- K-terminal reliability
- Monte-Carlo (MC) sampling
- Distance-constraint reliability – RHT sampling (Jin et. al., VLDB 2011)

# Baseline – MC Simulation + BFS




Uncertain Graph

MC Sampling + BFS



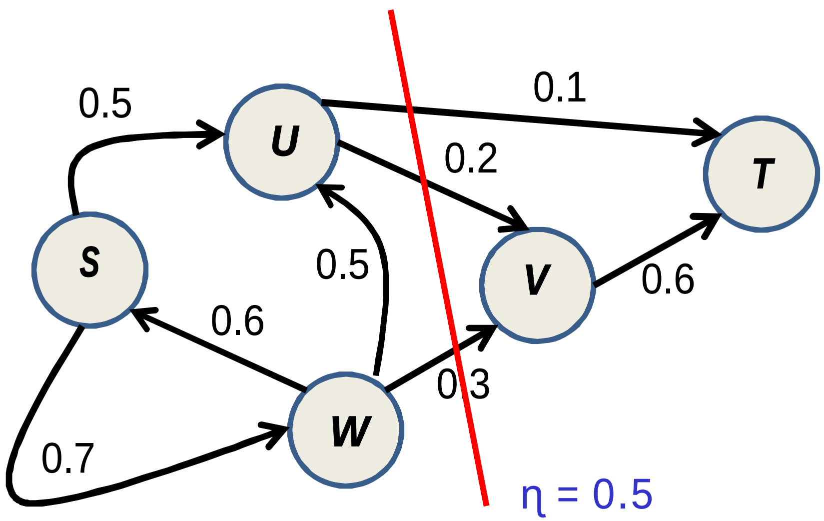

Certain Graph
(Possible World)

$$R(S,t) = \frac{1}{r} \sum_{G \sqsubseteq \mathcal{G}} P_G(S,t)$$

Number of Samples

# Can We Be More Efficient?

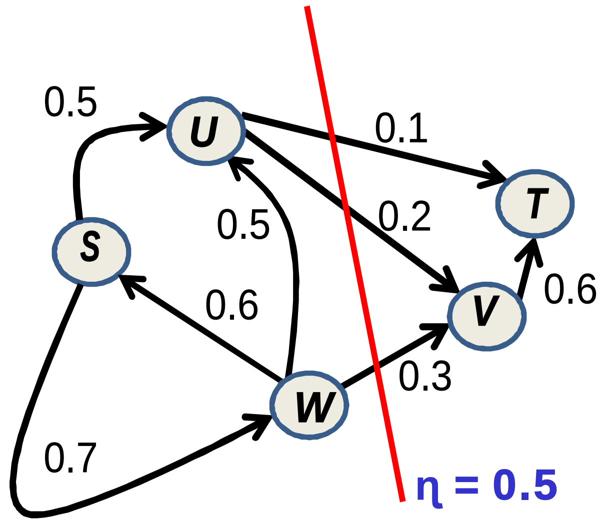
- Given a source node S and a probability threshold η ϵ (0, 1), can we quickly determine the nodes that are certainly not reachable from S with probability greater than or equal to η



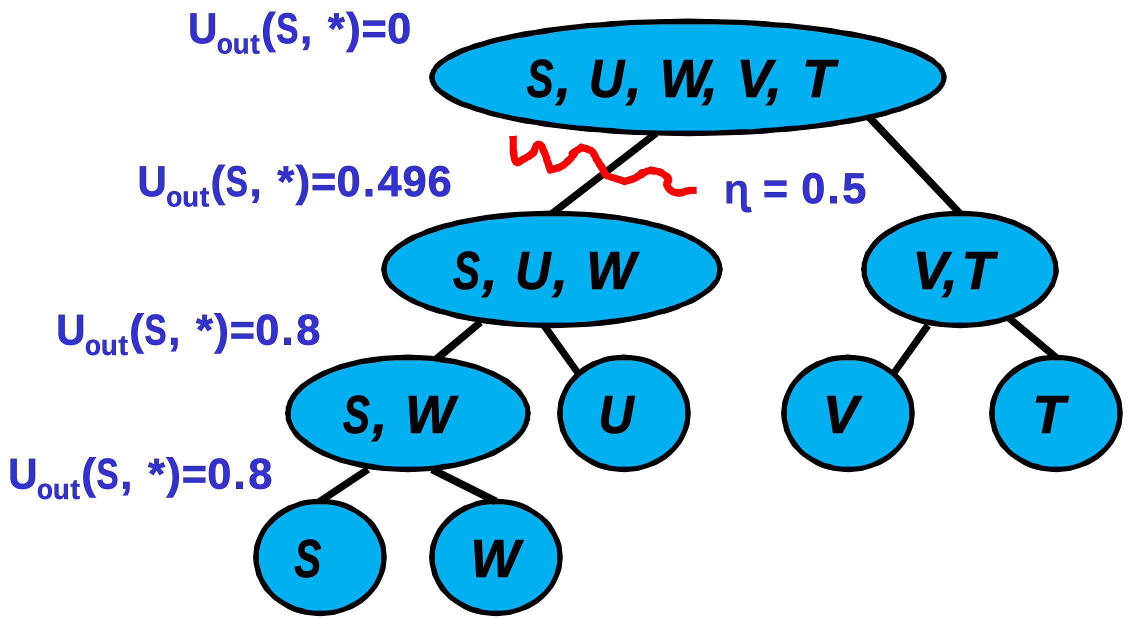

Uncertain Graph

- Indexing (offline)
- Filtering + Verification (Online)

# RQ-Tree Index

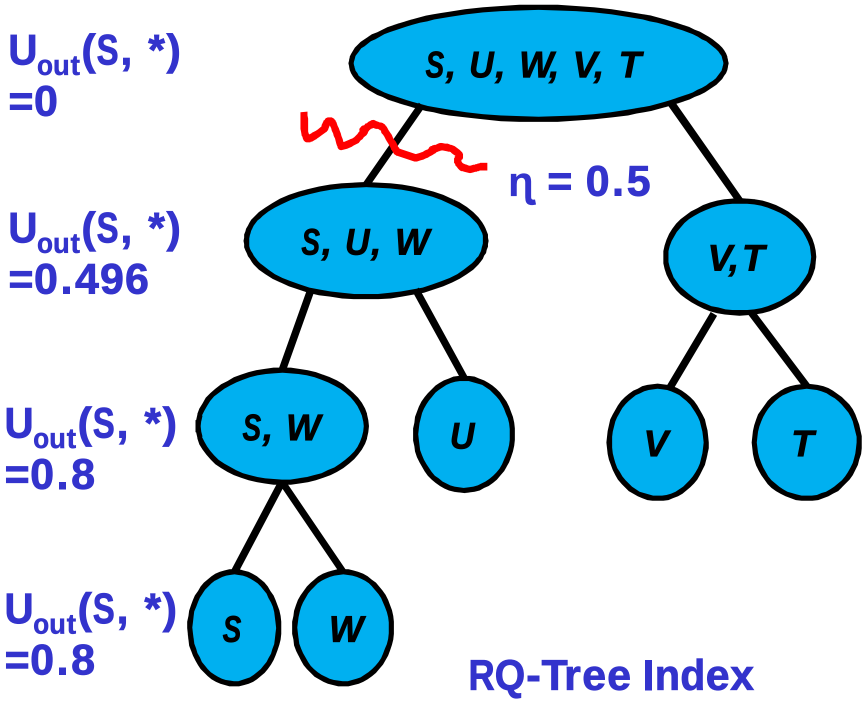



Uncertain Graph




RQ-Tree Index

# RQ-Tree: Filtering




**Max-Flow Min-Cut Based Upper Bound:**

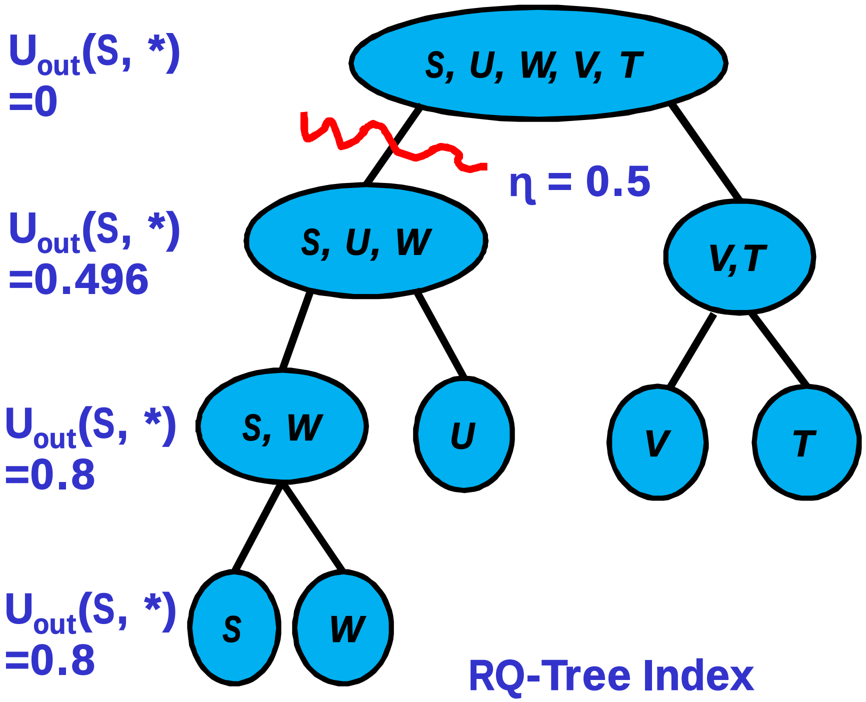
- Edge Capacity:

$$c(a) = -\log(1 - p(a))$$

- Compute Max-Flow f from S to Outside Cluster C

- $U_{out}(S, C) = 1 - \exp(-f)$

# RQ-Tree: Filtering




**Max-Flow Min-Cut Based Upper Bound:**

- Edge Capacity:

$$c(a) = -\log(1 - p(a))$$

- Compute Max-Flow f from S to Outside Cluster C
- $U_{out}(S, C) = 1 - \exp(-f)$

**Benefits:**

- No false negative (recall = 1)
- Computation limited only inside cluster C
- Incremental Max-Flow computation

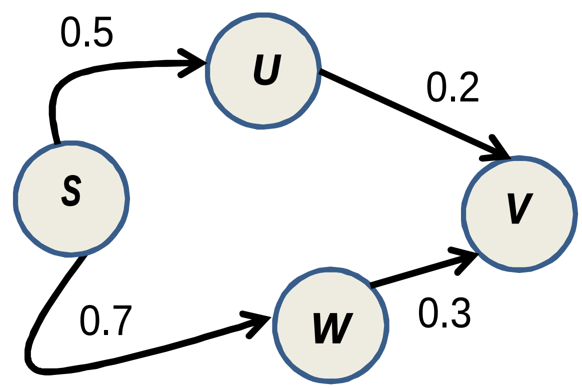
# RQ-Tree: Verification

**Sampling-based Verification:**

- MC-Sample + BFS over the sub-graph formed by the candidate set
- **Pros:** high precision, high recall
- **Cons:** verification could still be relatively expensive

**Lower-Bound-based Verification:**

- Most-Likely-Path
- **Pros:** precision = 1, high efficiency
- **Cons:** lower recall




**Pr(S-U-V) = 0.5 * 0.2 = 0.10**
**Pr(S-W-V) = 0.7 * 0.3 = 0.21**

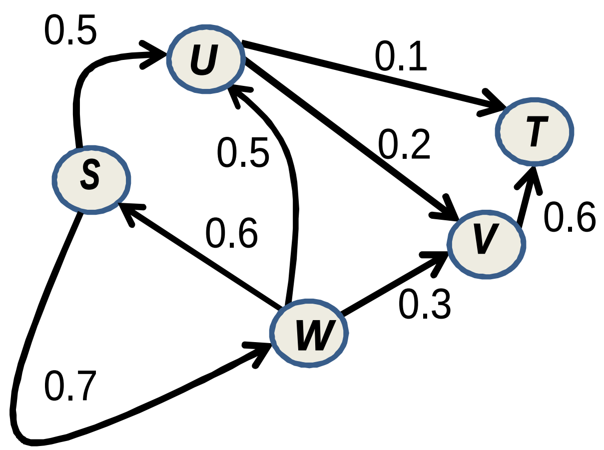
**Most-Likely-Path: (S-W-V)**

# RQ-Tree: Online Complexity

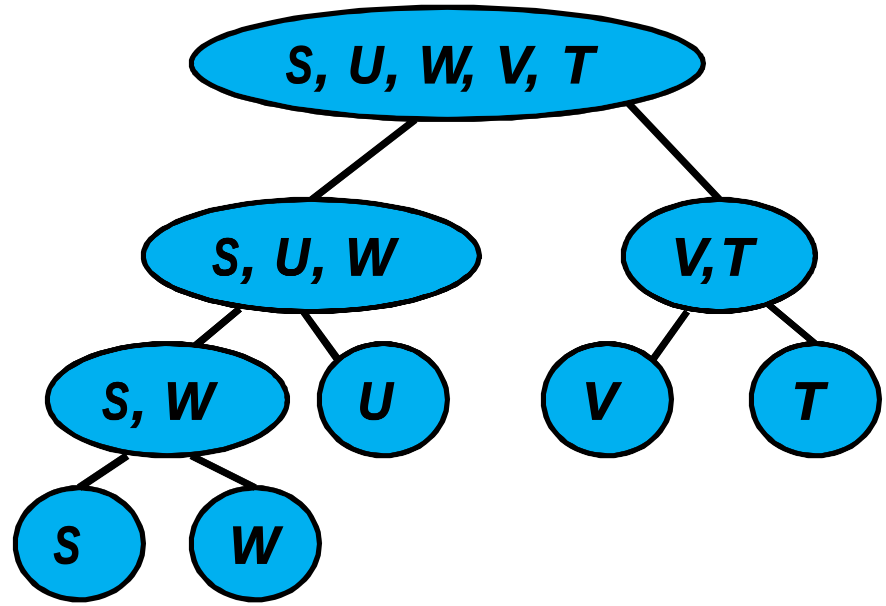
| MC Sampling | Recursive Sampling [VLDB '11] | RQ-Tree + MC-Sampling-based Verification [Our Method] | RQ-tree + Lower-Bound-based Verification [Our Method] |
|---|---|---|---|
| $O(K(m+n))$ | $O(n^2 d)$ | $O(\hat{m}\hat{n} + K(\hat{m} + \hat{n}))$ | $O(\hat{m}\hat{n})$ |

- K = No of Samples
- m = No of edges
- n = No of nodes
- $\hat{n}$ = No of nodes in the candidate set
- $\hat{m}$ = No of edges induced by the candidate nodes
- d = Diameter of the graph

# RQ-Tree Index Construction




Uncertain Graph




RQ-Tree Index

**Hierarchical Clustering:**

- Minimum-cut balanced bi-partition using METIS
- Edge weight:

$$w(a) = -\log(1 - p(a))$$

# Experimental Results

| | # Nodes | # Edges | #Arc Prob: Mean, SD, Quartiles |
|---|---|---|---|
| DBLP | 684 911 | 4 569 982 | 0.14 ± 0.11, {0.09, 0.09, 0.18} |
| Flickr | 78 322 | 20 343 018 | 0.09 ± 0.06, {0.06, 0.07, 0.09} |
| BioMine | 1 008 201 | 13 445 048 | 0.27 ± 0.21, {0.12, 0.22, 0.36} |

**Dataset Characteristics**

# Accuracy Results

| | RQ-Tree-MC | | | RQ-Tree-LB | | |
|---|---|---|---|---|---|---|
| | η=0.4 | η=0.6 | η=0.8 | η=0.4 | η=0.6 | η=0.8 |
| DBLP | 0.96 | 0.99 | 0.99 | 1 | 1 | 1 |
| Flickr | 0.97 | 0.98 | 0.98 | 1 | 1 | 1 |
| BioMine | 0.95 | 0.96 | 0.97 | 1 | 1 | 1 |

**Precision**

| | RQ-Tree-MC | | | RQ-Tree-LB | | |
|---|---|---|---|---|---|---|
| | η=0.4 | η=0.6 | η=0.8 | η=0.4 | η=0.6 | η=0.8 |
| DBLP | 0.99 | 0.99 | 1.00 | 0.75 | 0.87 | 0.91 |
| Flickr | 0.98 | 0.99 | 0.99 | 0.76 | 0.79 | 0.83 |
| BioMine | 0.97 | 0.98 | 0.98 | 0.77 | 0.81 | 0.85 |

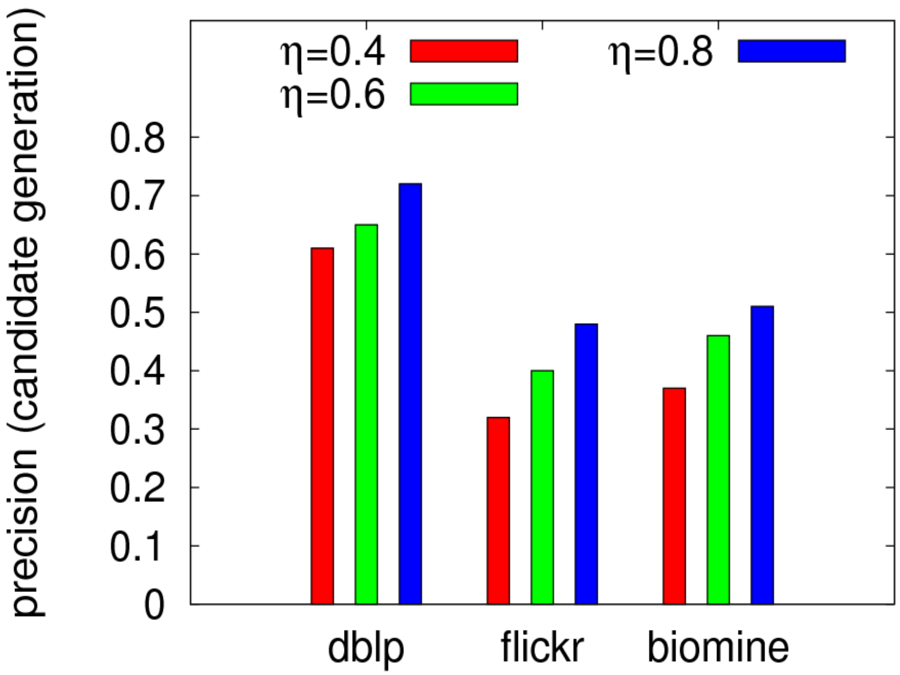
**Recall**

# Efficiency Results

| | RQ-Tree-MC | | | RQ-Tree-LB | | | MC |
|---|---|---|---|---|---|---|---|
| | η=0.4 | η=0.6 | η=0.8 | η=0.4 | η=0.6 | η=0.8 | All η |
| DBLP | 43 | 40 | 36 | 1.50 | 0.60 | 0.60 | 588 |
| Flickr | 60 | 59 | 55 | 0.21 | 0.20 | 0.17 | 114 |
| BioMine | 6062 | 5417 | 4974 | 1.00 | 0.50 | 0.50 | 25 608 |

**Online query-processing time (sec)**

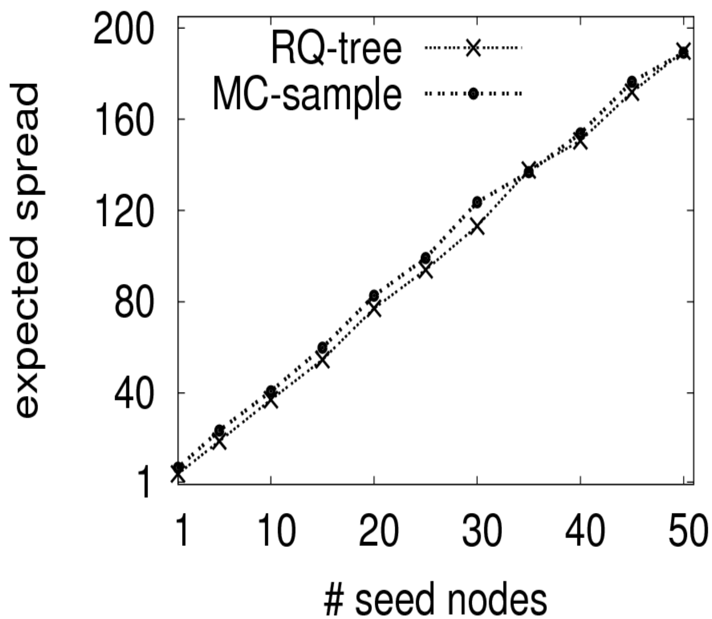
# Pruning Capacity of Filtering Phase



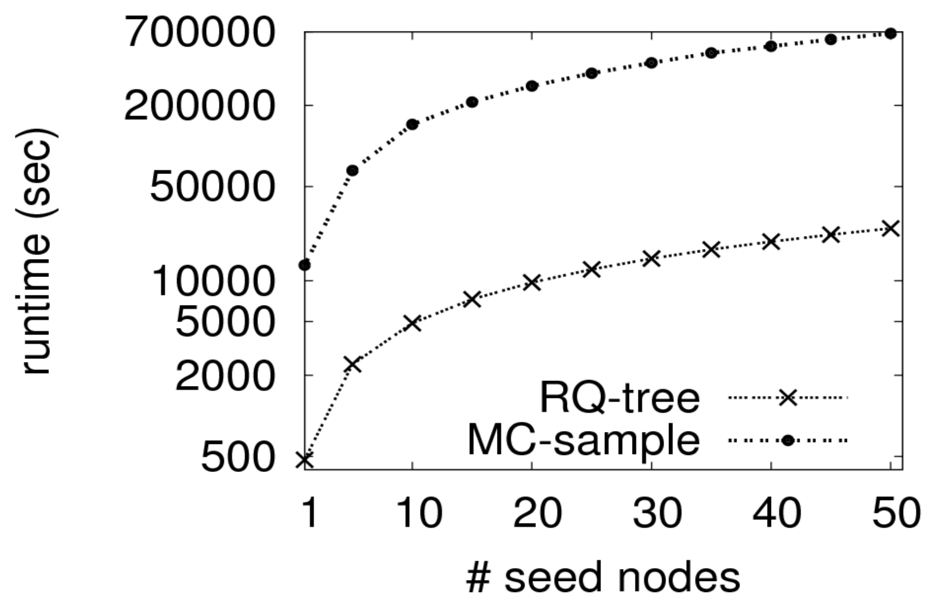

**Precision of Filtering Phase**

# RQ-Tree in Influence Maximization

- RQ-Tree index in multi-source reliability query and in influence maximization




**Expected Spread (Last.FM)**




**Top-k Seed Finding Time (Last.FM)**

# Conclusion

- Indexing method for answering online reliability queries efficiently and effectively.

- RQ-tree works very well with lower arc probabilities and with higher probability threshold.

- In future, we shall study reliability search queries when the arc probabilities are not independent.

# Questions?