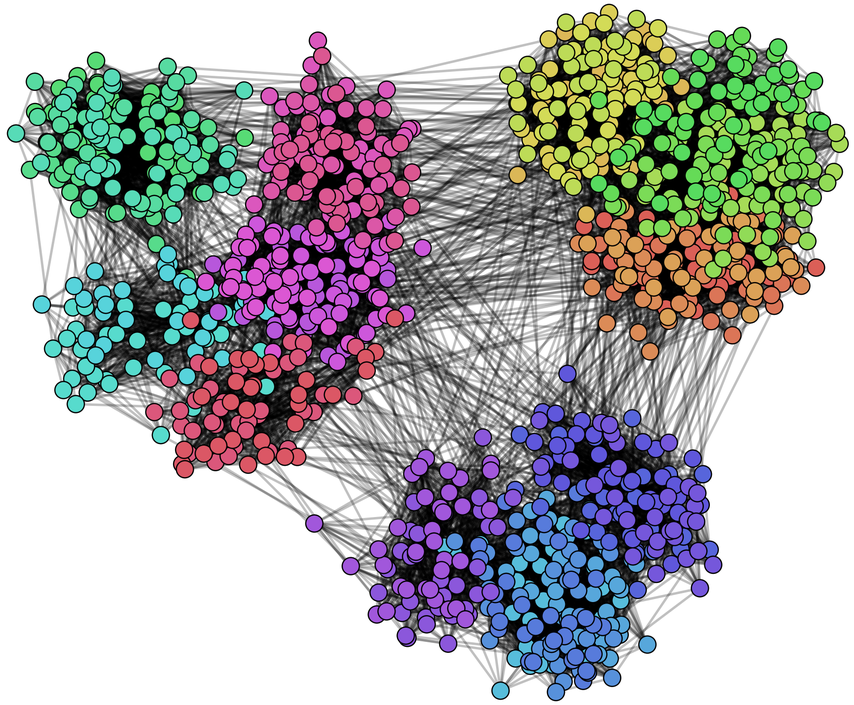
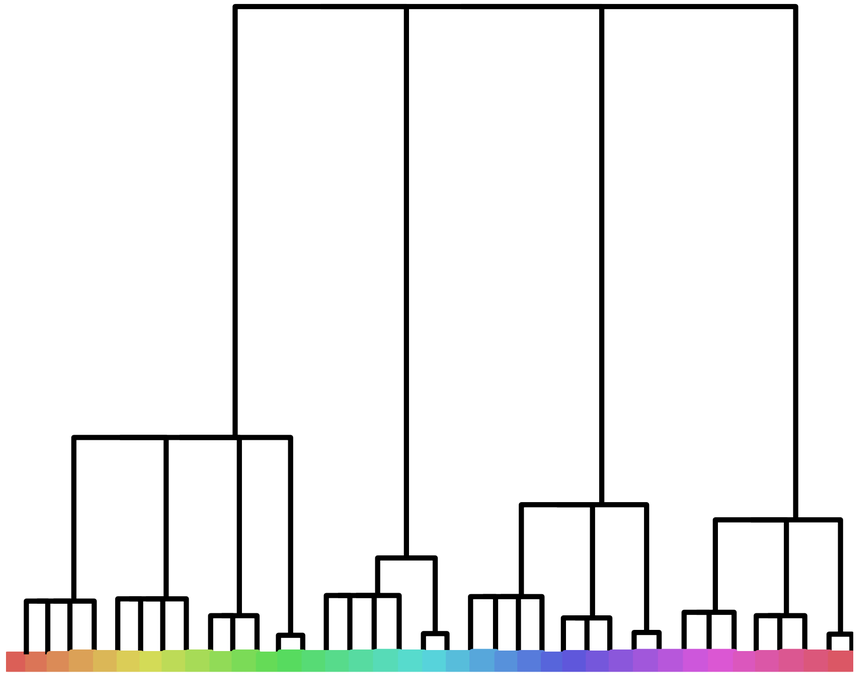
# Expected Probabilistic Hierarchies

Marcel Kollovieh, Bertrand Charpentier, Daniel Zügner, Stephan Günnemann

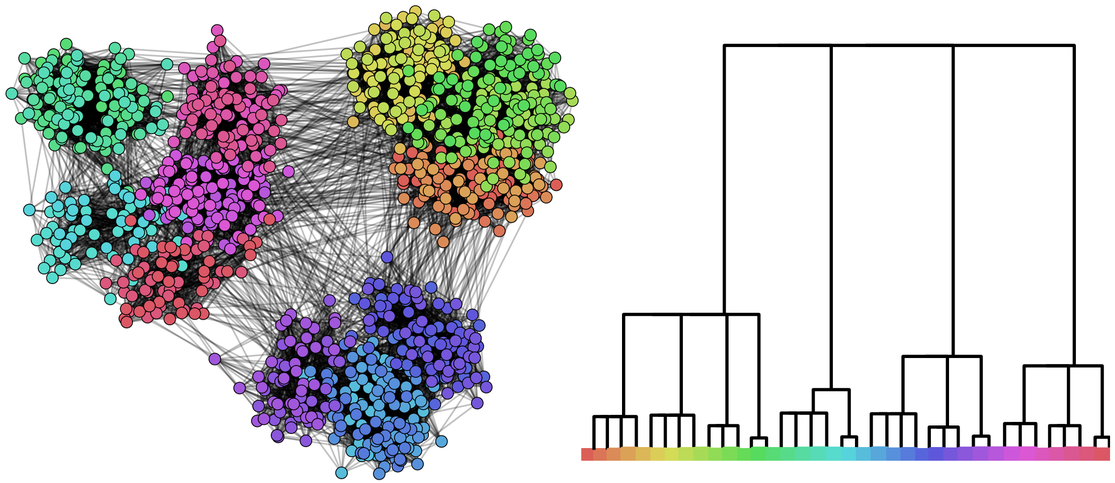
# Motivation





***Hierarchical clustering is essential to describe underlying structures***

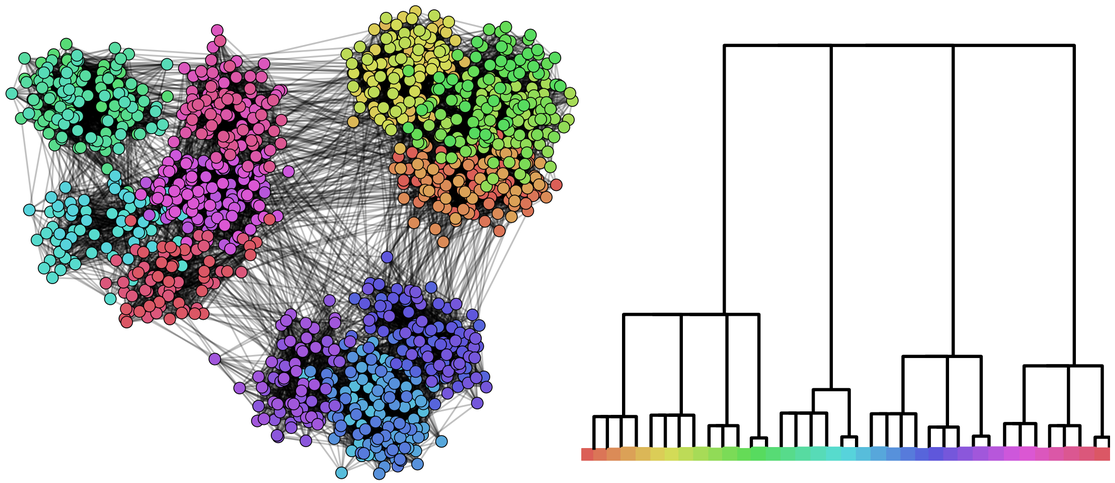
# Motivation



**Existing Hierarchical Clustering Algorithms:**

- Discrete approaches often based on heuristics and do not optimize global metrics
- Continuous relaxations do not necessarily align with their discrete counterparts

# Motivation



**Existing Hierarchical Clustering Algorithms:**

- Discrete approaches often based on heuristics and do not optimize global metrics
- Continuous relaxations do not necessarily align with their discrete counterparts

***How can we learn continuous hierarchies that align with the discrete problem?***

# Contributions

1. **Theoretical:** We analyze the theoretical properties of expected clustering scores and show that their optimal values are equal to their optimal discrete counterpart.

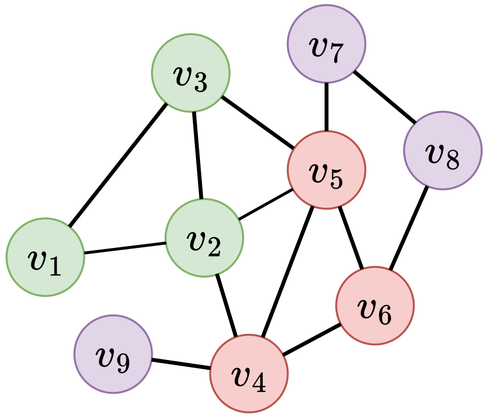
2. **Model:** We propose a new method called Expected Probabilistic Hierarchies (EPH) to optimize the expected scores using differentiable hierarchy sampling.

3. **Empirical:** In quantitative experiments, we show that EPH outperforms other baselines in 20/24 cases, including both graph and vector datasets. In qualitative experiments, we show that EPH provides meaningful hierarchies.

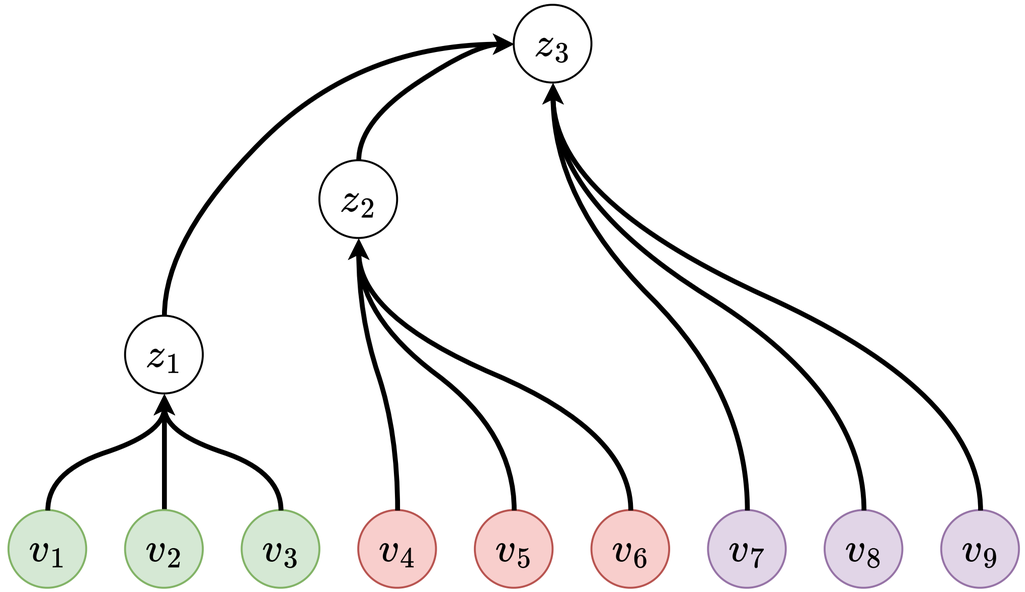
# Expected Probabilistic Hierarchies

## Problem

Graph




Hierarchy: $\hat{\mathcal{T}}$



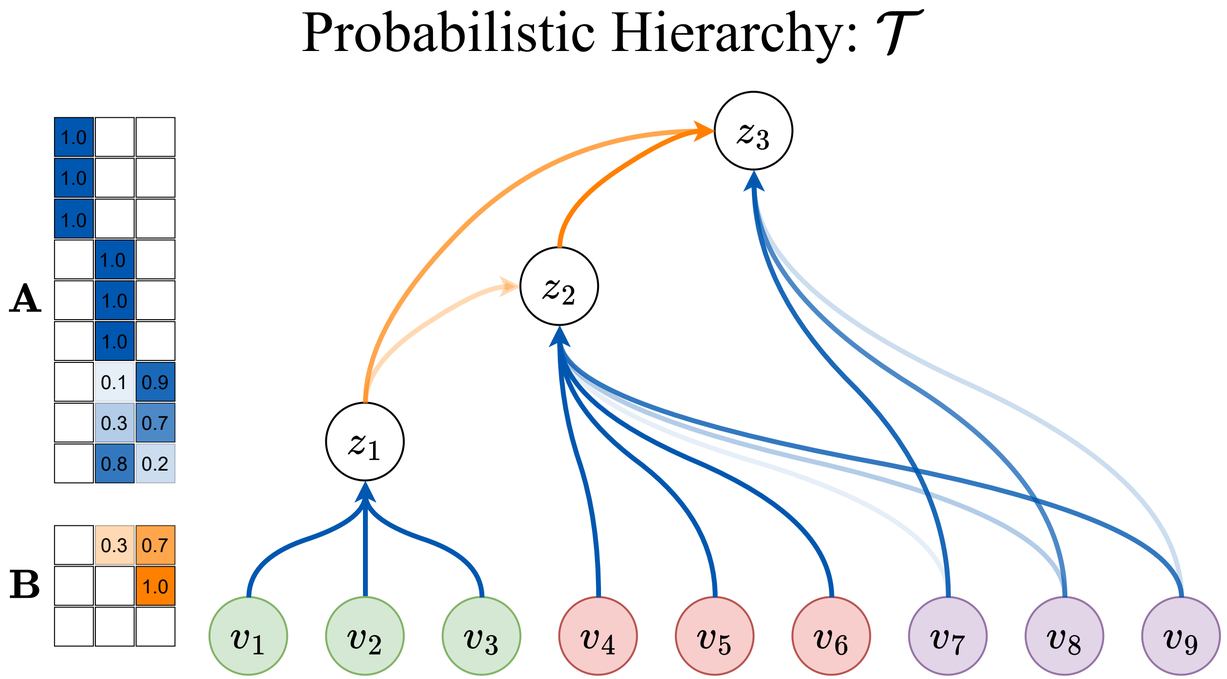
# Expected Probabilistic Hierarchies

## Probabilistic Hierarchies

*How can we parametrize continuous hierarchies?*

- **Probabilistic Hierarchies:**
  Relax outgoing edges onto the probability simplex

- **Tree-Sampling Procedure:**
  Sampling yields valid discrete hierarchies




**Zügner, Daniel, et al.** "End-to-end learning of probabilistic hierarchies on graphs." *ICLR* 2022

# Expected Probabilistic Hierarchies

## Expected Metrics

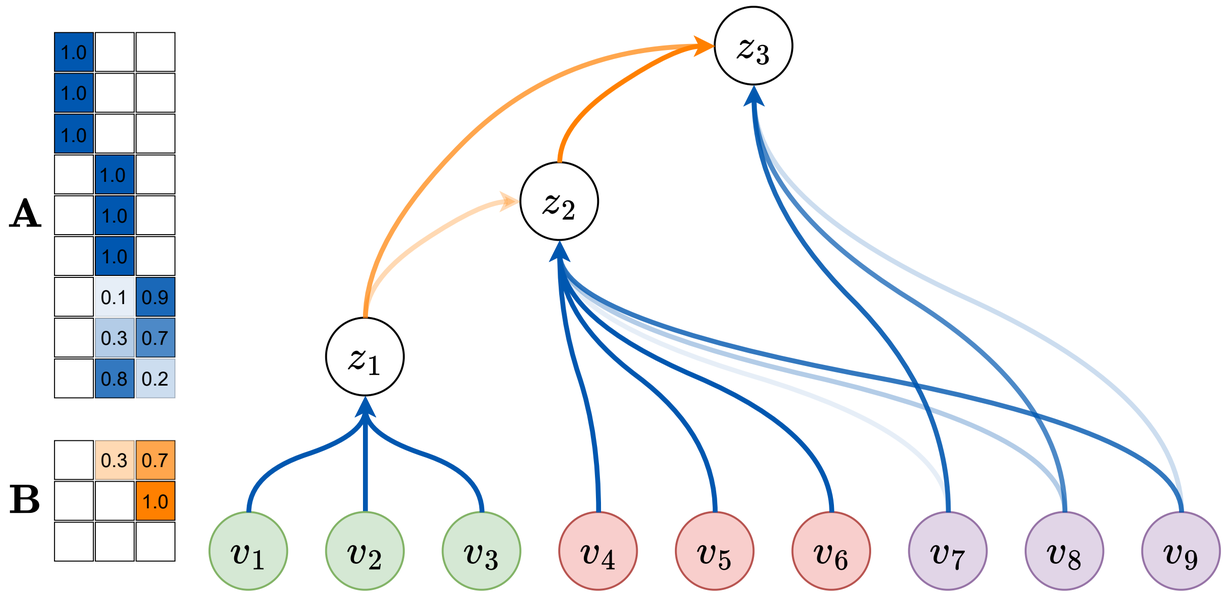
*How to quantify and learn the hierarchies?*

- **Dasgupta Cost: [1]**

$$\mathrm{Das}(\hat{\mathcal{T}}) = \sum_{\mathrm{v}_i, \mathrm{v}_j \in V} P(\mathrm{v}_i, \mathrm{v}_j) c(\mathrm{v}_i \wedge \mathrm{v}_j)$$

- **Tree-Sampling Divergence: [2]**

$$\mathrm{TSD}(\hat{\mathcal{T}}) = \mathrm{KL}(p(\mathrm{z}) \| q(\mathrm{z}))$$





[1] **Dasgupta, Sanjoy.** "A cost function for similarity-based hierarchical clustering." 2016

[2] **Charpentier, Bertrand, and Thomas Bonald.** "Tree sampling divergence: an information-theoretic metric for hierarchical graph clustering." 2019

# Expected Probabilistic Hierarchies

## Expected Metrics

*How to quantify and learn the hierarchies?*

- **Dasgupta Cost:**

$$\mathrm{Das}(\hat{\mathcal{T}}) = \sum_{\mathrm{v}_i, \mathrm{v}_j \in V} P(\mathrm{v}_i, \mathrm{v}_j) c(\mathrm{v}_i \wedge \mathrm{v}_j)$$

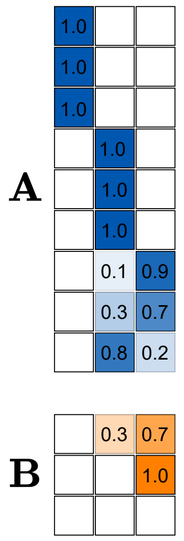
- **Tree-Sampling Divergence:**

$$\mathrm{TSD}(\hat{\mathcal{T}}) = \mathrm{KL}(p(\mathrm{z}) \| q(\mathrm{z}))$$

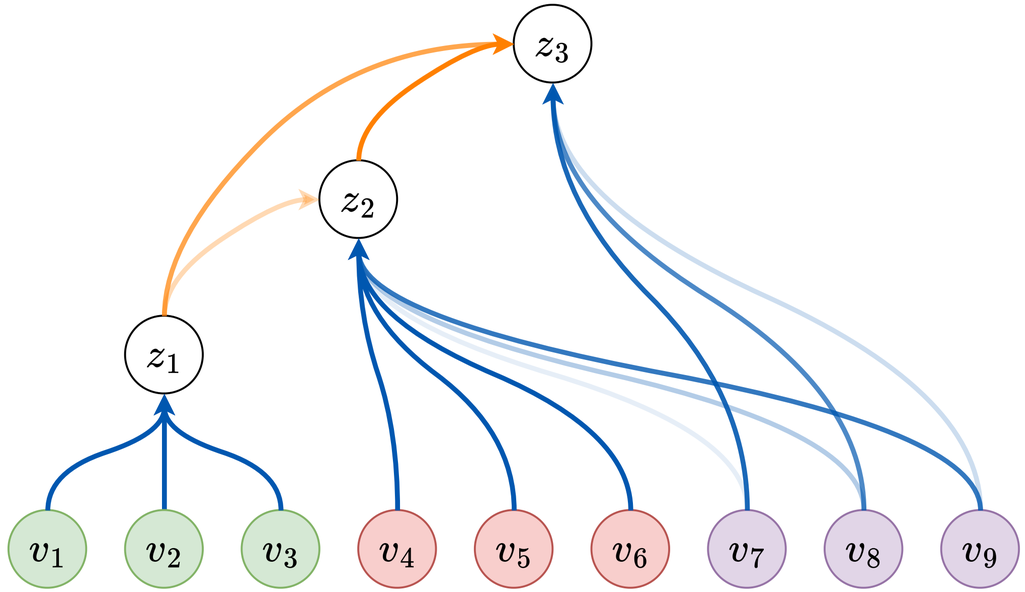
*Train using expected metrics:*

$$\min_{\mathbf{A},\mathbf{B}} \mathbb{E}_{\hat{\mathcal{T}}}\left[\mathrm{Das}(\hat{\mathcal{T}})\right] \quad \text{and} \quad \max_{\mathbf{A},\mathbf{B}} \mathbb{E}_{\hat{\mathcal{T}}}\left[\mathrm{TSD}(\hat{\mathcal{T}})\right]$$

Approximated using Monte Carlo and Gumbel-Softmax trick

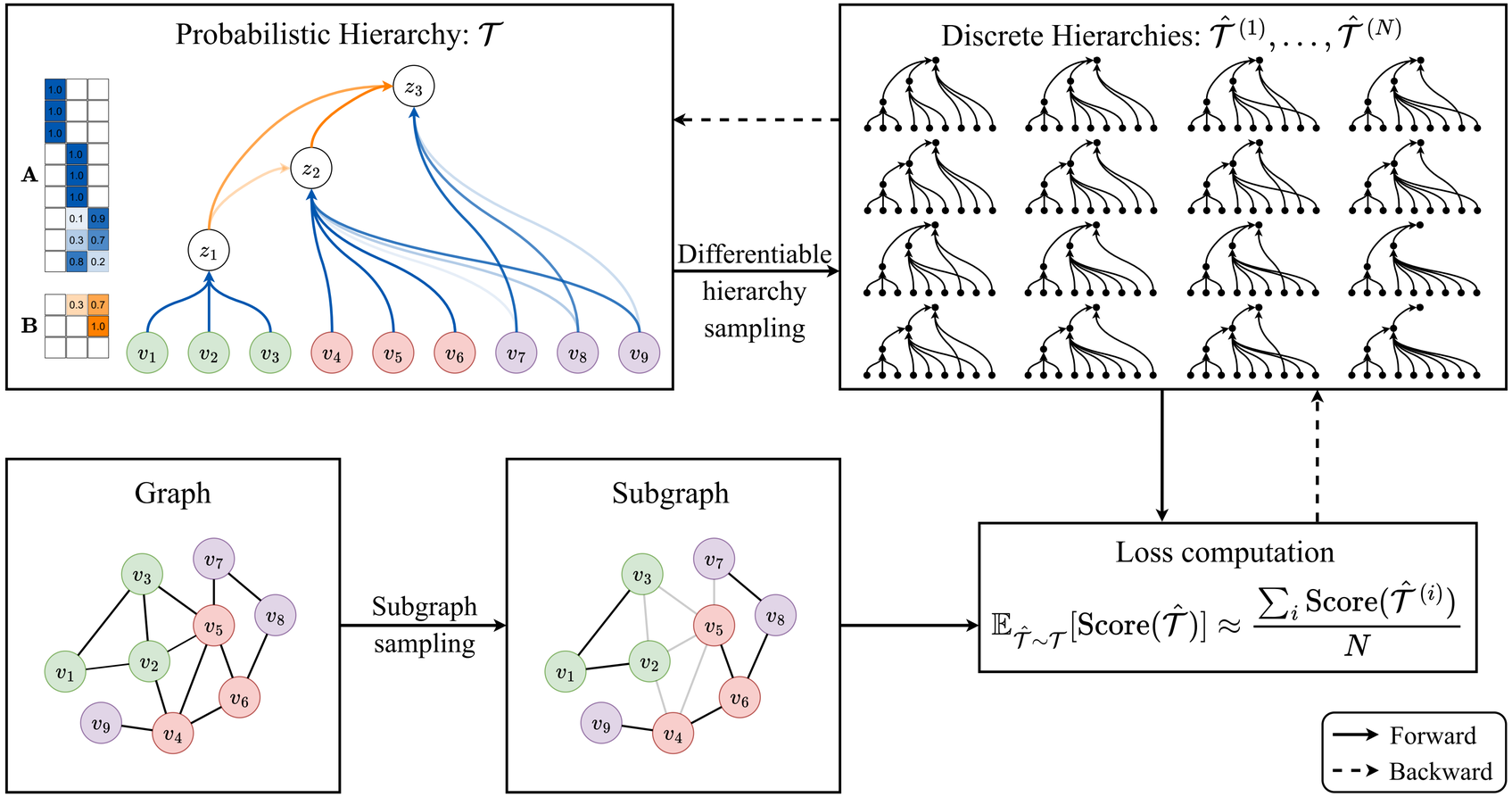



Probabilistic Hierarchy: $\mathcal{T}$



# Expected Probabilistic Hierarchies

## Overview



# Results

## Graph Datasets

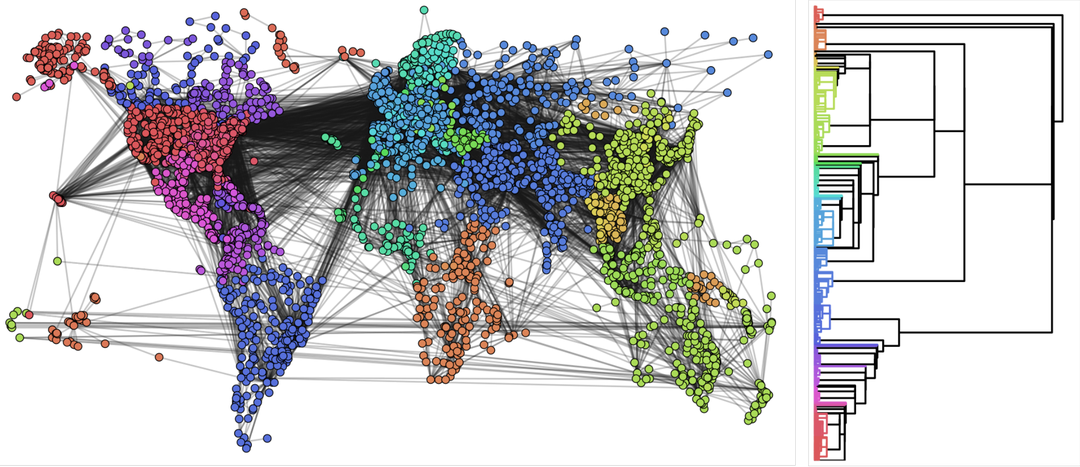
| | Dasgupta cost (↓) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | PolBl. | Brain | Cites. | Genes | Cora-ML | OpenF. | WikiP. | DBLP |
| WL | 338.52 | 567.90 | 137.80 | 270.18 | 301.68 | 379.68 | 660.12 | OOM |
| AL | 355.61 | 556.68 | 83.69 | 196.50 | 292.77 | 363.40 | 658.04 | 36,463 |
| Louv. | 344.47 | 582.45 | 158.79 | 247.27 | 335.57 | 501.29 | 798.75 | 40,726 |
| RSC | 307.70 | 526.17 | 85.41 | 188.82 | 264.62 | 367.36 | 630.53 | OOM |
| UF | 331.79 | 508.30 | 91.86 | 208.51 | 305.43 | 410.17 | 560.45 | OOM |
| gHHC | 349.71 | 595.70 | 147.17 | 308.42 | 313.29 | 390.21 | 672.84 | 87,344 |
| HypHC | 272.81 | 519.96 | 416.38 | 632.02 | 594.23 | 529.04 | 678.45 | OOM |
| FPH | 238.65 | 425.70 | 76.03 | 182.91 | 257.42 | 355.61 | 482.40 | 31,687 |
| EPH | **235.50** | **400.20** | **74.01** | **176.57** | **238.28** | **312.31** | **456.26** | **30,600** |

| | Tree-sampling divergence (↑) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | PolBl. | Brain | Cites. | Genes | Cora-ML | OpenF. | WikiP. | DBLP |
| WL | 26.59 | 25.13 | 62.14 | 60.93 | 52.76 | 50.59 | 42.18 | OOM |
| AL | 25.25 | 28.91 | 67.80 | 66.72 | 55.30 | 52.02 | 43.15 | 38.99 |
| Louv. | 28.86 | 30.74 | 68.09 | 67.51 | 58.18 | 52.97 | 47.01 | 41.40 |
| RSC | 28.04 | 29.19 | 67.39 | 66.28 | 56.14 | 52.01 | 44.86 | OOM |
| UF | 21.77 | 24.49 | 60.13 | 59.45 | 48.42 | 47.64 | 42.37 | OOM |
| gHHC | 24.70 | 25.62 | 59.53 | 54.20 | 49.56 | 51.36 | 41.08 | 16.29 |
| HypHC | 19.65 | 7.26 | 18.98 | 13.00 | 19.18 | 26.82 | 23.92 | OOM |
| FPH | 31.37 | 32.75 | **69.38** | **67.78** | **59.55** | 57.58 | 49.87 | 41.62 |
| EPH | **32.05** | **34.24** | 69.36 | 67.75 | 59.41 | **57.83** | **50.23** | **42.74** |

## Vector Datasets

| | Dasgupta cost (↓) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dataset | Zoo | Iris | Glass | Digits | Segm. | Spam. | Letter | Cifar |
| WL | 56.28 | 69.98 | 122.16 | 1126.77 | 1266.17 | 2962.62 | 12241 | 32979 |
| AL | 56.31 | 69.48 | 121.64 | 1121.68 | 1258.22 | 2952.21 | 12181 | 32972 |
| SL | 57.67 | 70.71 | 126.33 | 1166.05 | 1368.21 | 3042.28 | 13166 | 33314 |
| CL | 55.78 | 70.10 | 123.02 | 1140.26 | 1277.48 | 2971.59 | 12396 | 33131 |
| Louv. | 56.26 | 72.31 | 125.94 | 1126.48 | 1238.35 | 2916.48 | 11946 | 32940 |
| RSC | 55.94 | **69.10** | 121.45 | 1119.43 | 1237.20 | 2917.07 | 11895 | **32907** |
| UF | 56.28 | 69.40 | 122.43 | 1137.53 | 1322.86 | 2998.17 | 13090 | OOM |
| gHHC | 60.09 | 69.63 | 123.33 | 1119.74 | 1269.28 | 3018.44 | 12151 | 33089 |
| HypHC | 56.05 | 69.22 | 121.52 | 1118.08 | 1233.07 | 2921.38 | 11930 | OOM |
| FPH | 56.13 | 69.13 | 122.00 | 1132.84 | 1238.45 | 2933.56 | 12197 | 33224 |
| EPH | **55.77** | **69.10** | **120.94** | **1117.58** | **1230.60** | **2916.17** | **11894** | 32913 |

## Qualitative



# Takeaways

- **Expected Probabilistic Hierarchies:**
  Probabilistic framework optimizing expected scores
- **Theoretical Properties:**
  Continuous global optima align with their discrete counterparts, unlike previous objectives
- **Empirical Results:**
  EPH outperforms other approaches on graph and vector datasets in 20/24 cases

Full Paper