Structural Clustering of Large-Scale Graph Databases

Professor Dr. Petra Mutzel
based on work with Till Schäfer

Lehrstuhl für Algorithm Engineering

Fakultät für Informatik

TU Dortmund

Selected Topics on Combinatorial Optimization

Vienna Graduate School on Computational Optimization

Lectures 6, November 12, 2018

Outline

Literature

- Till Schäfer, Petra Mutzel: StruClus: Structural Clustering of Large-Scale Graph Databases, in: 13th International Conference on Advanced Data Mining and Applications (ADMA) 2017, 343–359 (Spotlight Paper)
- M.A. Hasan, V. Vhaoji, S. Salem, J. Besson, and M.J. Zaki: ORIGAMI: mining representative orthogonal graph patterns, in: 7th IEEE International Conference oN Data Mining (ICDM) 2007, 153–162

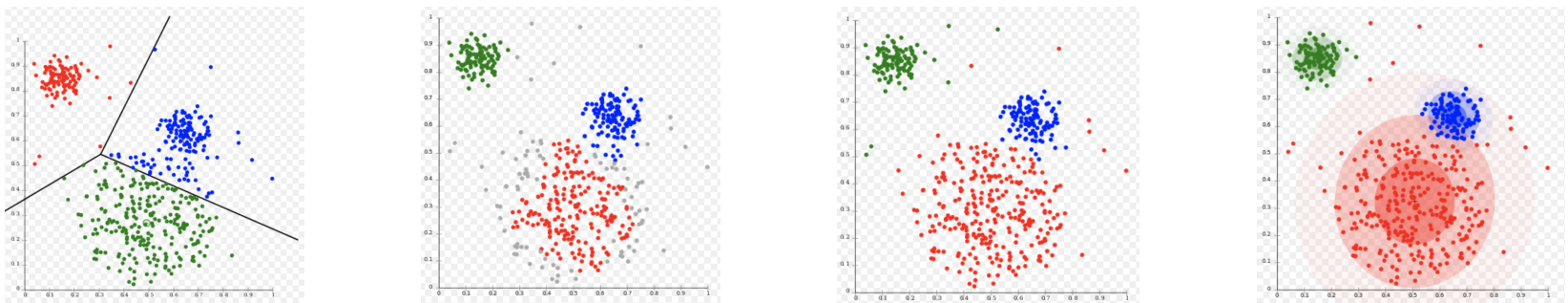
Data Clustering

Definition (Clustering Problem)

Clustering is the task of grouping a set of objects such that objects in the same group (cluster) are more similar to each other than objects in different groups (separate).

Remarks

- Many different models (e.g., overlapping, hierarchical)
- E.g., distance-based methods in vector space
- E.g., centroid-based clustering, like k-means clustering



Source: `https://en.wikipedia.org/wiki/K-means_clustering`

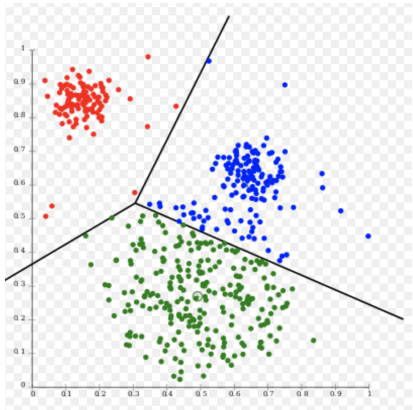
K-Means Clustering

Definition (K-Means Clustering)

Find k cluster centers and assign all the objects to the nearest cluster center, so that the sum of the squared distances from the cluster are minimized.

k-Means Algorithm

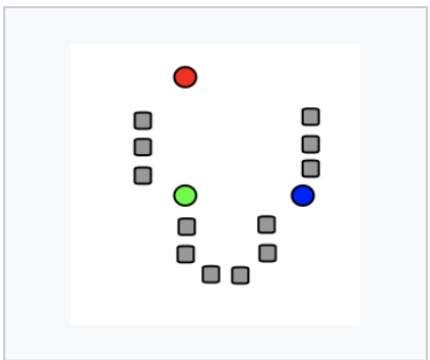
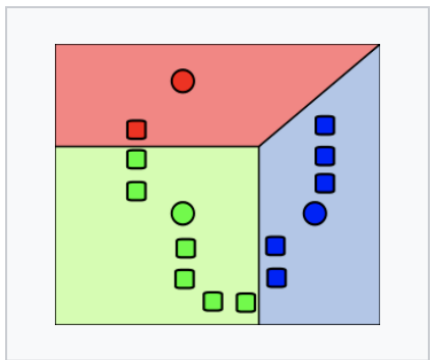
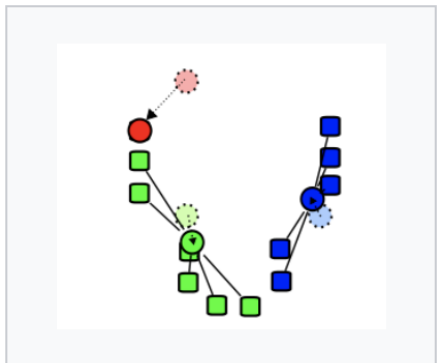
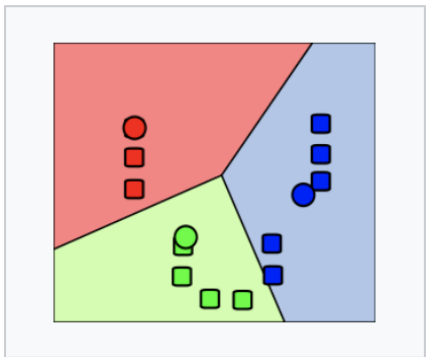
- Decision Problem is NP-complete
- k-Means Algorithms (Lloyd) finds local optima



Source: `https://en.wikipedia.org/wiki/K-means_clustering`

K-Means Algorithm

1: **function** kMeans(k)
2: Initialisation: randomly choose k centroids
3: **while** not convergent **do**
4: update centroids
5: assign all points to its closest centroid
6: **end while**
7: **end function**

Source: https://en.wikipedia.org/wiki/K-means_clustering

Evaluation Methods if Ground Truth is Known

Let $D = \{d_1, \ldots, d_{|D|}\}$ be the true number of classes of N elements, and $C = \{c_1, \ldots, c_{|C|}\}$ be the number of clusters found by an algorithm.

Definition (Purity)

For each cluster $c \in C$: count the number of data points from the most common class in D (ground truth). Purity is then defined of the average of these values for all clusters:

$$\frac{1}{N} \sum_{c \in C} \max_{d \in D} |c \cap d|$$

- Purity good measure for homogenieity (every cluster contains only elements from a single class)
- Purity does not measure completeness (all elements of each class are assigned to the same cluster)

Evaluation Methods if Ground Truth is Known

Definition (Fowlkes-Mallows Index)

The Fowlkes-Mallows Index calculates the similarity between the clusters and the classes. Let TP denote the number of point pairs that are present in the same cluster and the same class (true positives). Similarly define the number of false positives (FP) and false negatives (FN). Then:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

Definition (Normalized Variation of Information)

$$VI(D, C) = \frac{H(D|C) + H(C|D)}{H(D)},$$

where $H(C)$ (resp. $H(C|D)$) are (conditional) entropies ($H(c) \neq 0$.

Structural Graph Clustering

Clustering Graphs

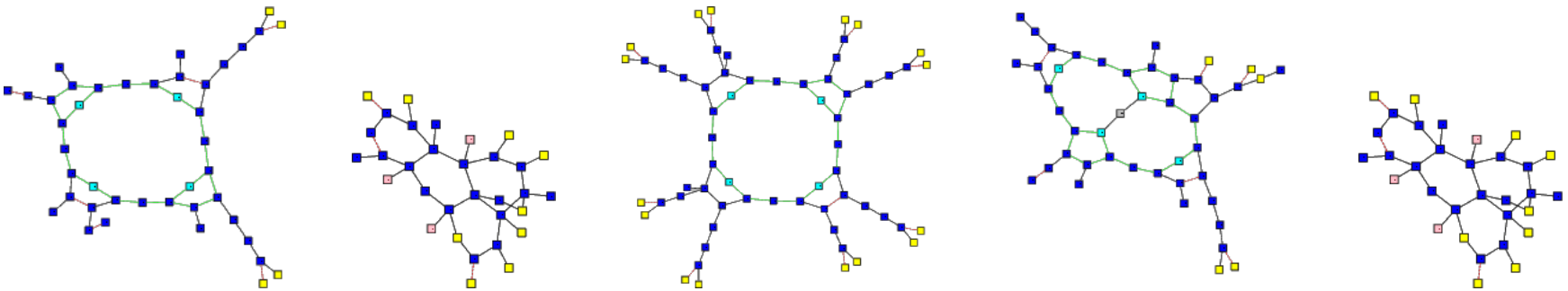
Given: Set of labeled graphs $\mathcal{X} = \{G_1, \ldots, G_n\}$

Task: Find a partition of $\mathcal{X}$ that

- maximizes cluster homogeneity
- has well separated clusters

Setting

- Limited graph size $|G|$
- Large dataset size $|\mathcal{X}|$ $(\gg 10^6)$



Structural Graph Clustering

Clustering Graphs

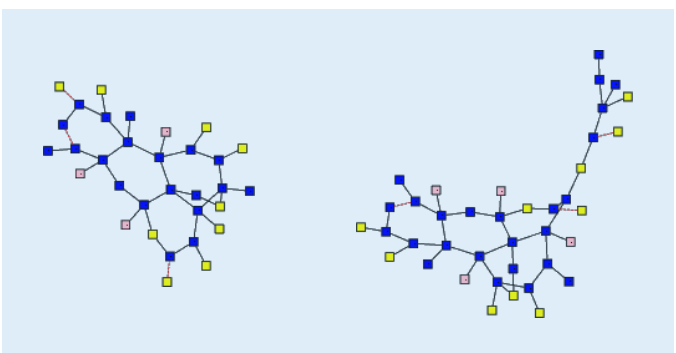
Given: Set of labeled graphs $\mathcal{X} = \{G_1, \ldots, G_n\}$

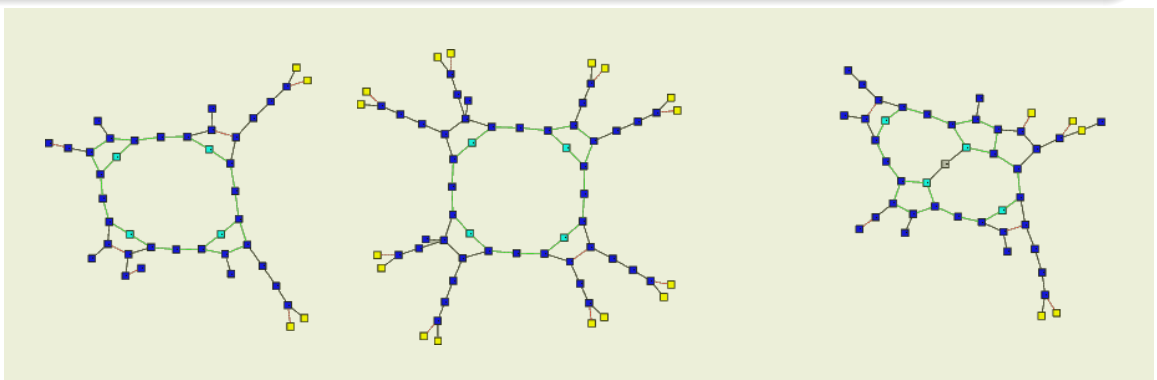
Task: Find a partition of $\mathcal{X}$ that

- maximizes cluster homogeneity
- has well separated clusters

Setting

- Limited graph size $|G|$
- Large dataset size $|\mathcal{X}|$ $(\gg 10^6)$





Clustering Graphs - View on Data

Distance

Data $\Leftrightarrow$ Pairwise distances

Vector Space

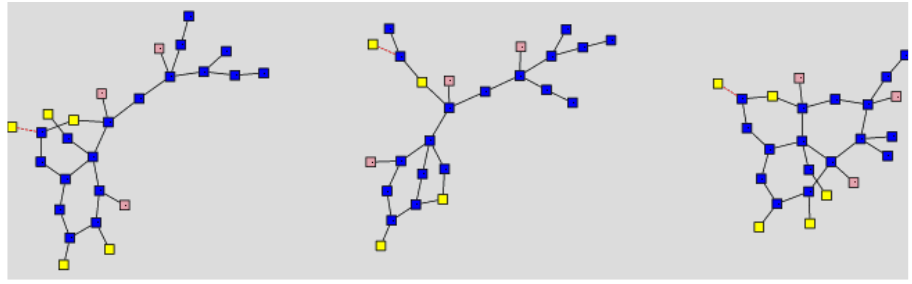
Data $\Leftrightarrow$ Set of vectors

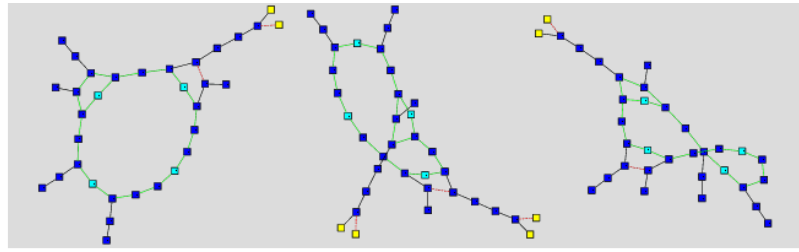
Problems with Generalization

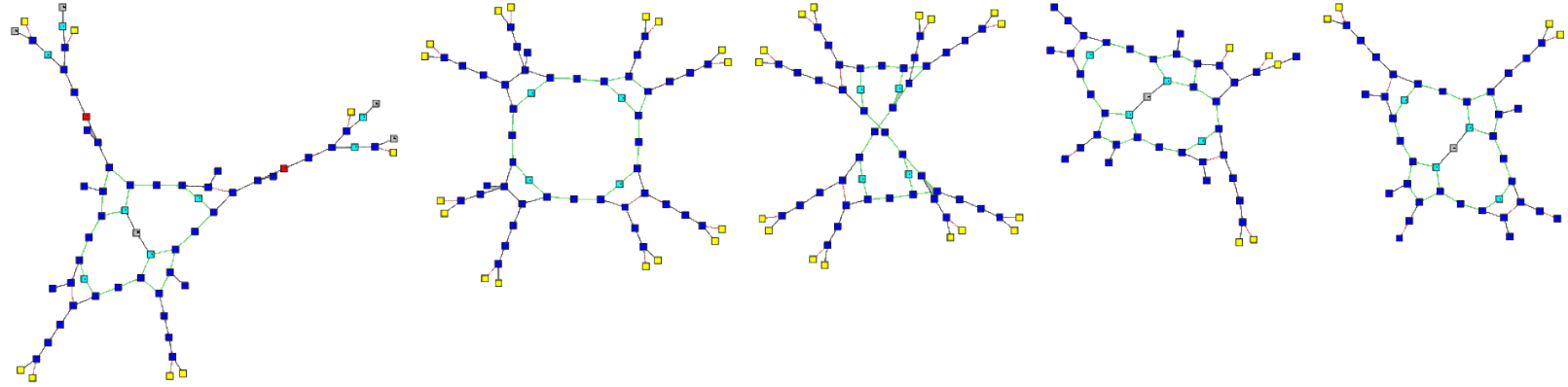
- High (intrinsic) dimensionality $\rightarrow$ Concentration effect
- Lossy transformation
- Low interpretability

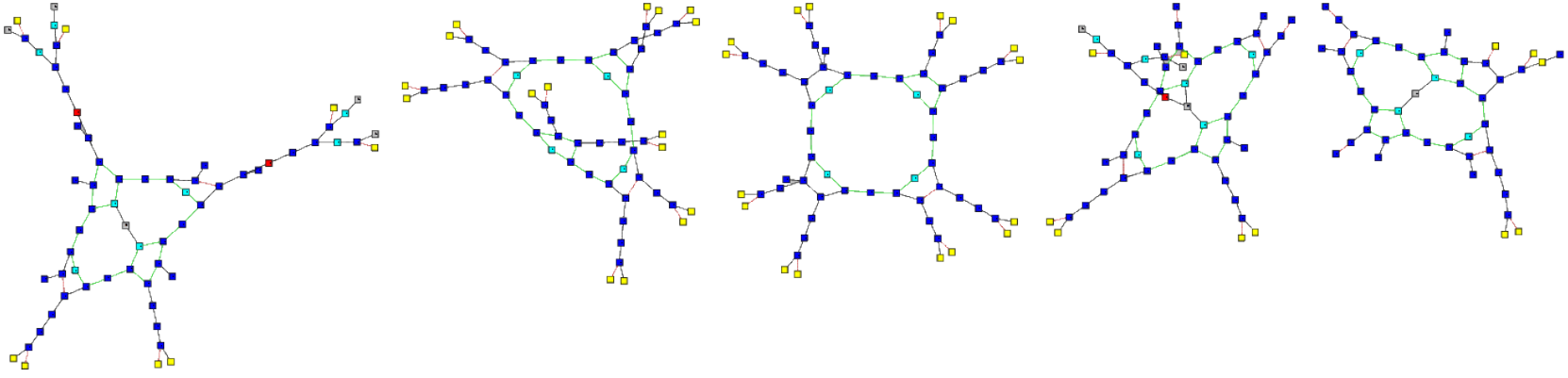
Structural Graph Clustering

Data $\Leftrightarrow$ Set of graphs



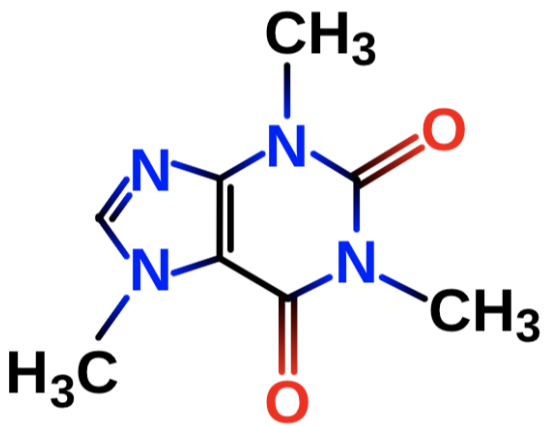


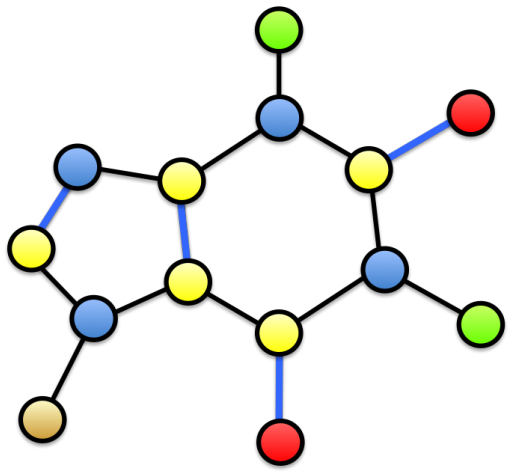




Motivation: Rational Drug Design

- Which molecules are active against disease X?
- Which molecules have a similar function/effect? (Reduction of side effects)
- Which molecules may have an increased effectiveness?
- High-throughput screening for promising candidates





- Molecules can be modelled as graphs with attributes
- Direct relationship between structure and effects

$\rightarrow$ Graph similarity

CHIPMUNK: A virtual synthesizable molecule library

Goal

Create ligand molecule database that

- is accessible,
- leads to novel drug discoveries (uniqueness of library).

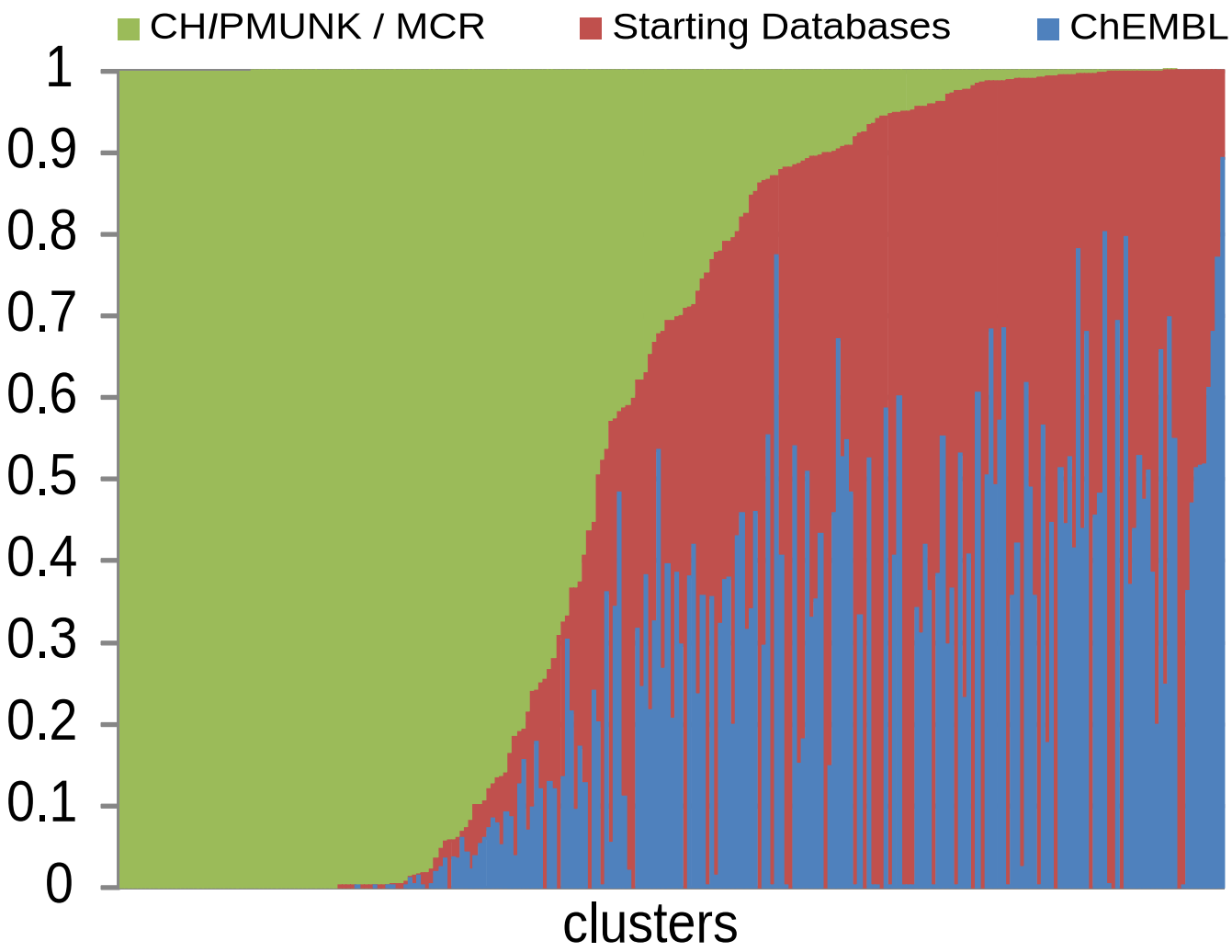
Generation

- purchasable building blocks (ZINC, eMol $\cup$ MolPort)
- simulating chemical reactions that are easy to reproduce in the wet lab (MCR, MedChem, Heterocycle)
- overall $\approx$ 95 million molecules

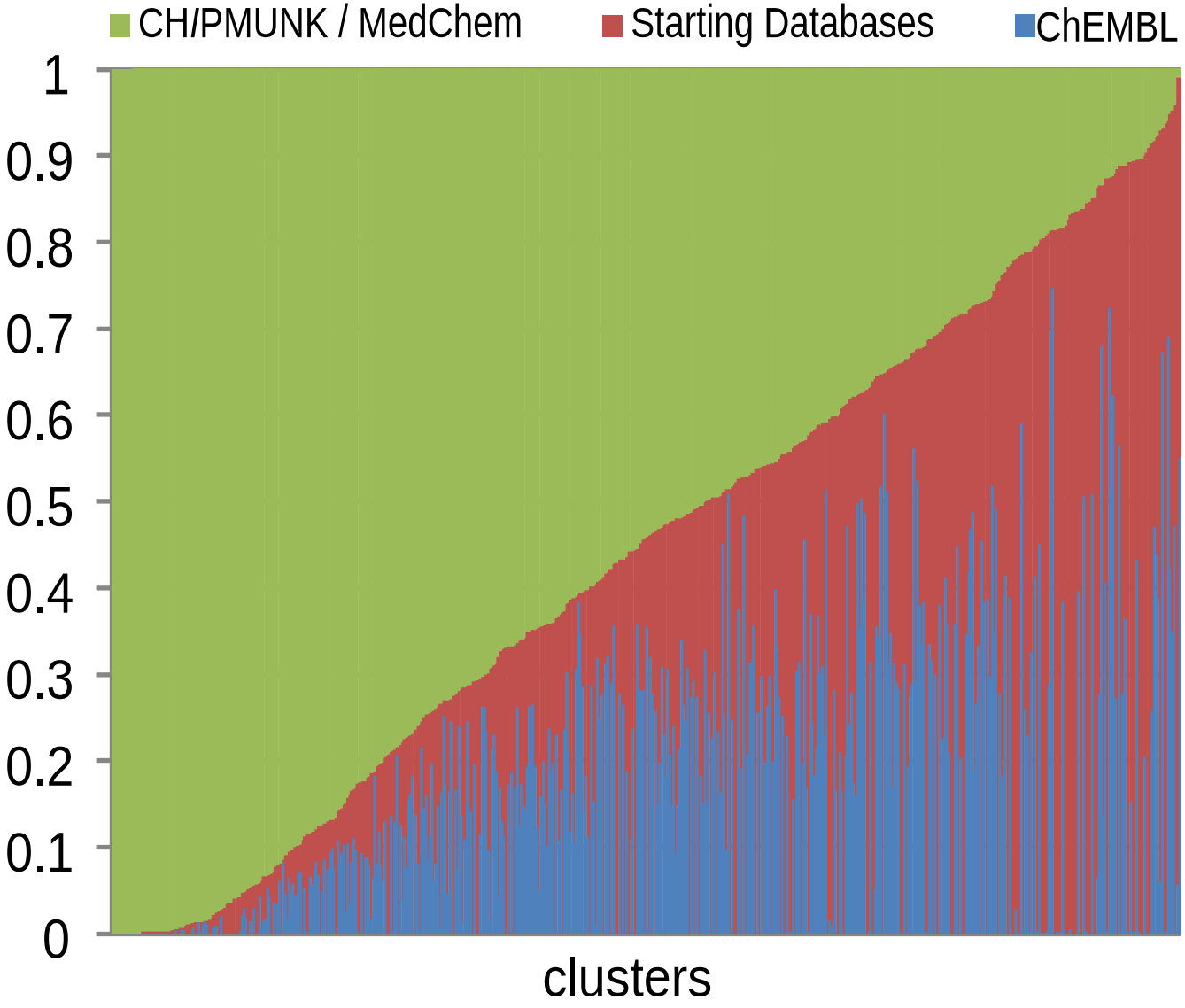
Uniqueness Evaluation

Test Setup

- Cluster CH/PMUNK ∪ ZINC ∪ eMol ∪ MolPort ∪ ChEMBL
- using StruClus
- inspect database distribution per cluster
- up to 67 mio. of molecules (MedChem)
- If a library covers a unique subspace, clusters with low portions of the other databases will be visible.

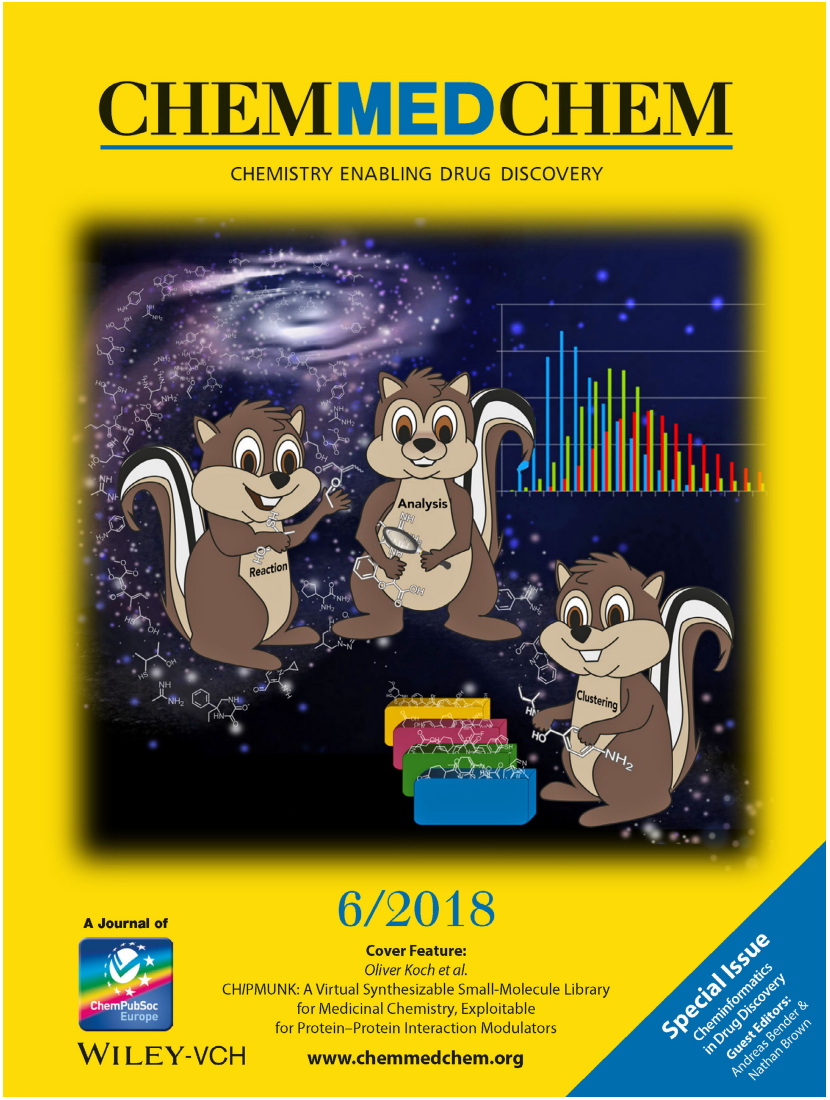


y-axis: relative share of structures for MCR library



y-axis: relative share of structures for MedChem

Data Analysis: Collect, Analyse, Evaluate



Humbeck, Weigang, Schäfer, Mutzel, Koch: CHIPMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, ChemMedChem 2018 + Cover Feature

The StruClus Algorithm (ADMA 2017)

Novelties of StruClus

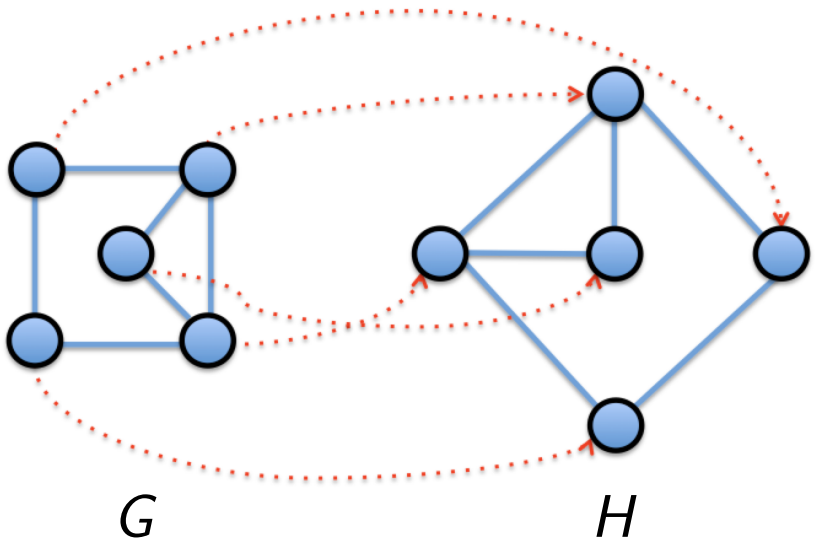
- Structural clustering algorithm for general labeled graphs that scales to very large datasets
- Intuitive cluster description using representatives
- Similarity based on subgraph isomorphism of representatives
- Active selection of representatives based on homogeneity and separation constraints
- Error bounded support sampling
- Balanced per-cluster coverage with dynamically adjusted minimum support

Graph Similarity

Definition (Graph Isomorphism)

Let $G = (V_G, E_G)$ and $H = (V_H, E_H)$ be simple graphs. A bijective mapping $\pi : V_G \to V_H$ is called graph isomorphism if the following holds:

$$\forall\, v, w \in V_G : (v, w) \in E_G \iff (\pi(v), \pi(w)) \in E_H$$



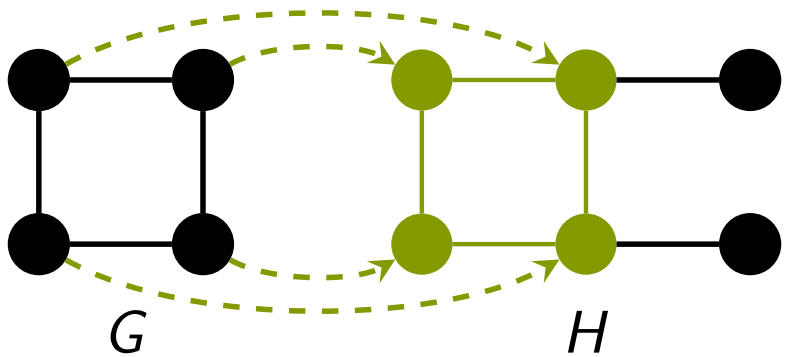
Two graphs are called isomorph ($G_1 \simeq G_2$), if a graph isomorphism exists.

Subgraph Isomorphism

Definition (Subgraph Isomorphism)

Let $G = (V_G, E_G)$ und $H = (V_H, E_H)$
An injection $\psi : V_G \to V_H$ is a subgraph isomorphism from G to H if

$$\forall u, v \in V(G) : (u, v) \in E(G) \Rightarrow (\psi(u), \psi(v)) \in E(H)$$



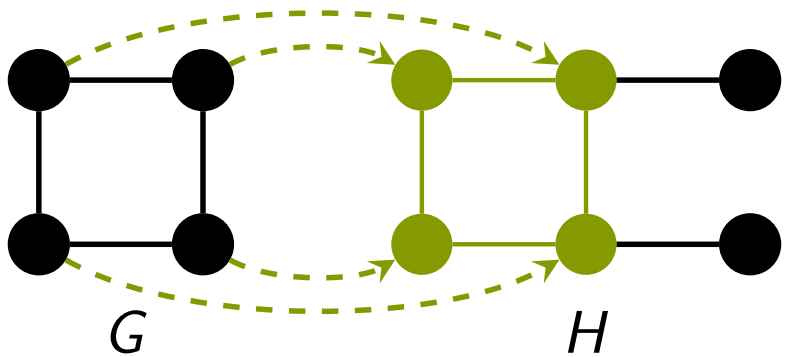
Notation: $G \subseteq H$

Complexity: Decision problem is NP-complete

Frequent Subgraphs

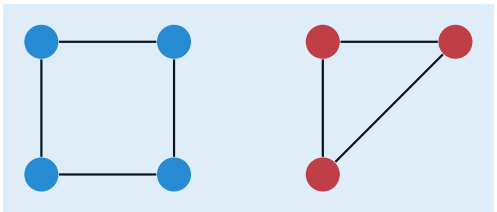
Definition

- The support $\sup(G)$ of a graph G over a set of graphs $\mathcal{G}$ is the fraction of graphs in $\mathcal{G}$ for which G is subgraph isomorphic.
- G is frequent if its support is at least the minimum support threshold $sup_{\min}$.
- A frequent subgraph is maximal if there exists no frequent supergraph of G.
- The coverage of a graph H by a graph $G \subseteq H$ is defined as $cov(G, H) = \frac{|G|}{|H|}$

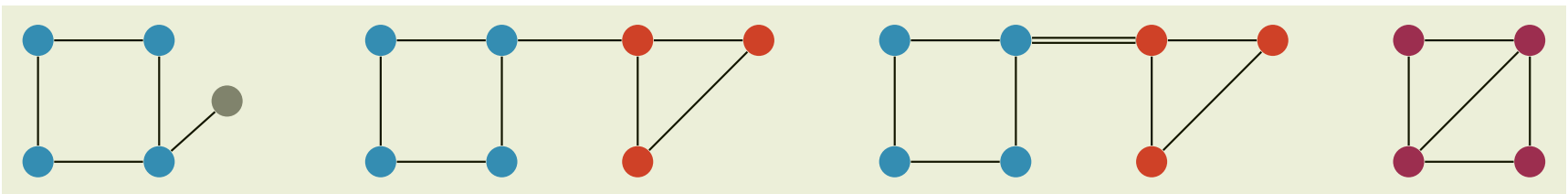


Projected Structural Cluster

Representatives



Members



Invariant: For each member there exists a subgraph isomorphism from at least one representative.

Our Contribution: StruClus Overview

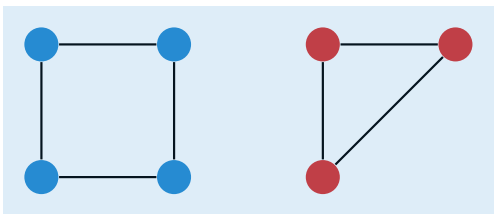
1: **function** StruClus
2: apply pre-clustering via maximal frequent subgraphs
3: **while** not convergent **do**
4: split clusters
5: merge clusters
6: update representatives
7: assign graphs to closest cluster
8: **end while**
9: **end function**

Representative Selection

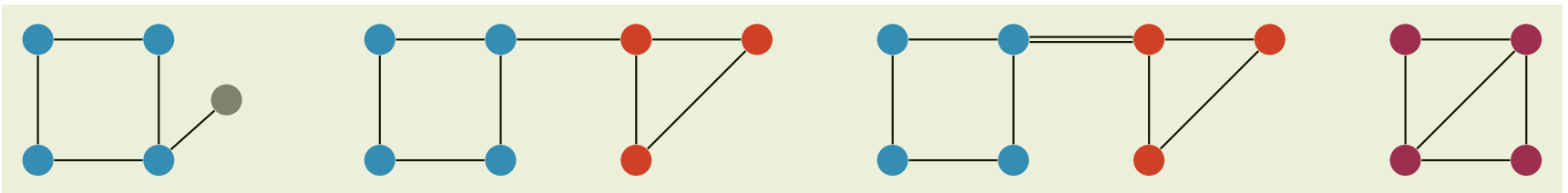
Criteria

- High support
  - Support: Fraction of subgraph isomorphic graphs
- High coverage
- Discriminative

Representatives



Members

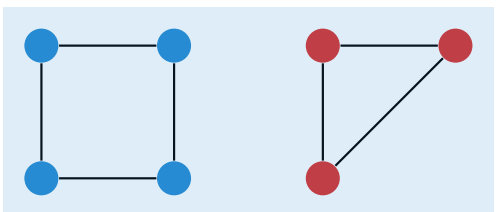


Representative Selection

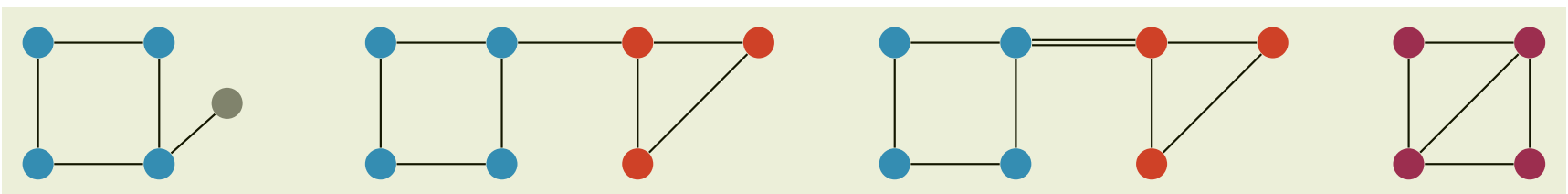
Two Stage Process

- Candidate generation
- Representative Selection

Representatives



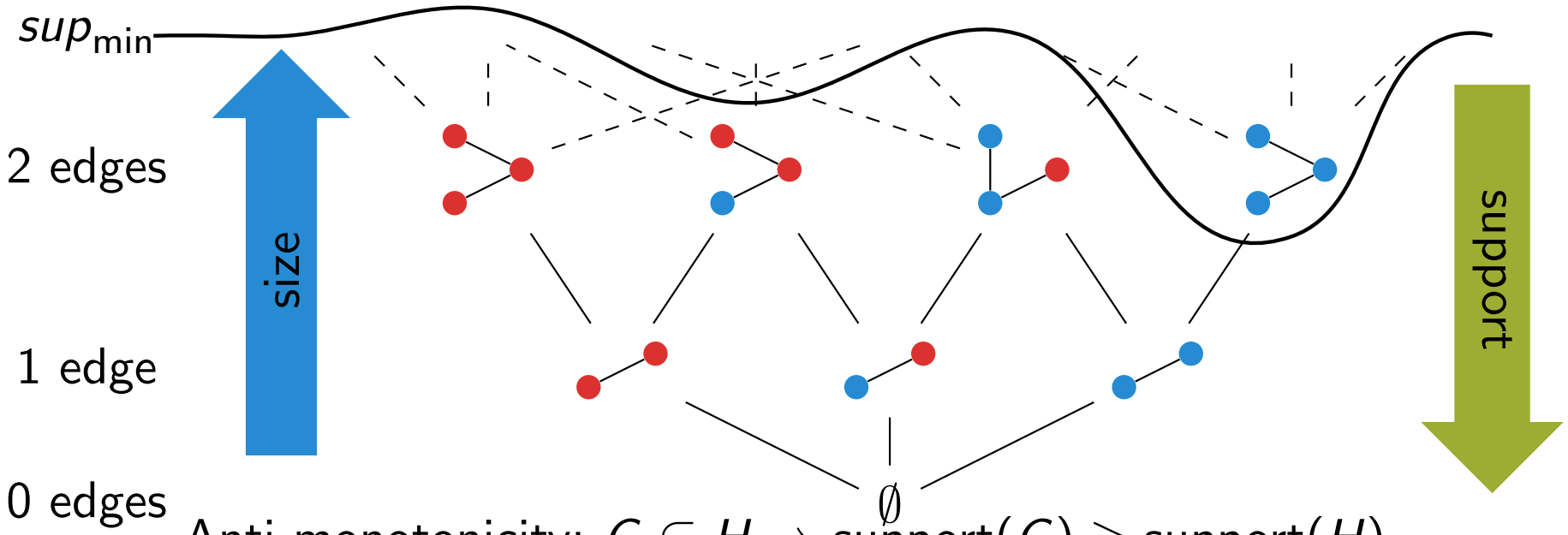
Members



Representative Candidate Generation

Frequent Subgraphs

Graph F is frequent $\Leftrightarrow$ support$(F) \geq sup_{\text{min}}$



Anti-monotonicity: $G \subseteq H \Rightarrow$ support$(G) \geq$ support(H)

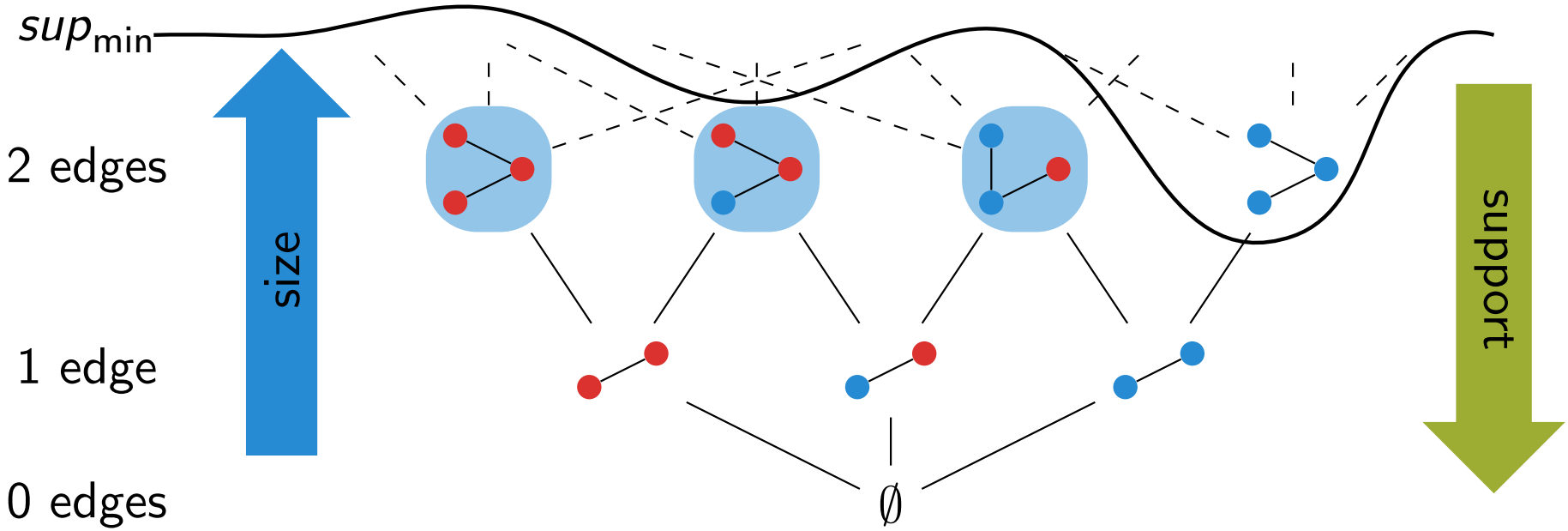
Observations

- Support vs. Coverage

Representative Candidate Generation

Maximal Frequent Subgraphs

Graph F is frequent $\Leftrightarrow$ support$(F) \geq sup_{\text{min}}$



Anti-monotonicity: $G \subseteq H \Rightarrow$ support$(G) \geq$ support(H)

Observations

- Support vs. Coverage
- Balancing with sup_{min}

Sampling Maximal Frequent Subgraphs

Random Walk on Lattice [Hasan et. al.]

- start with a random frequent vertex
- add frequent edges until maximal

Observation

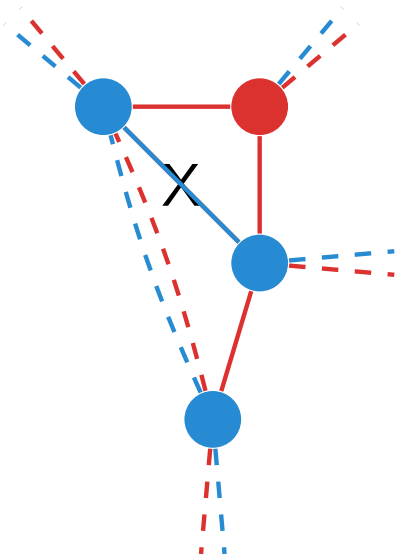
- For each mined subgraph G_{cand}

$$\left(\binom{V_{\text{cand}}}{2} + V_{\text{cand}}\right) \; |\text{FE}(\mathcal{G})| \; |\mathcal{G}|$$

subgraph isomorphism tests (WC).

- $\text{FE}(\mathcal{G})$: frequent edges of $\mathcal{G}$
- $\rightarrow$ major bottleneck for StruClus

Example



Our Contribution: Support Sampling

Support Sampling with Controlled Error Rate α

Mine a single maximal frequent subgraph with probability $\geq 1 - \alpha$

For each support test:

- Start with sample $\mathcal{H}_{\text{min}} \subseteq \mathcal{G}$
- Binomial test: below / above sup_{min}
- Test fails? $\rightarrow$ Double sample size

Support Sampling with Controlled Error Rate

Proposition: Multiple Hypothesis Testing Correction

Let $\mathcal{G}$ be a set of undirected labeled graphs, $|\mathcal{H}_{\text{min}}|$ the minimal sample size, $\text{FE}(\mathcal{G})$ the set of all frequent paths of length one, sup_{min} the minimum support threshold, and V_{max} the $(1 - sup_{\text{min}})$-quantile of the sorted graph sizes in $\mathcal{G}$.
Then the maximal number of binomial tests to construct a maximal frequent substructure over $\mathcal{G}$ is bounded by:

$$\left\lceil \log \frac{|\mathcal{G}|}{|\mathcal{H}_{\text{min}}|} \right\rceil \left(\binom{V_{\text{max}}}{2} + V_{\text{max}} \right) |\text{FE}(\mathcal{G})|$$

Representative Selection

Properties of good representatives

- A good representative explains a large portion of its cluster.
- It should be supported by a large fraction of C
- It should cover a large fraction of each graph
- It should be supported only by small fraction of other clusters (discriminative)

Ranking function for selection

For dataset X, cluster c and representative R, we select the highest ranked sampled subgraphs as representatives:

$$rank(R) = \frac{|c_R||R|}{\sum_{G \in c_R} |G|}(\sup(R, c) - \sup(R, X)),$$

where $c_R = \{G \in c | R \subseteq G\}$

Reassignment of Graphs to Cluster

- Each graph G is assigned to its most similar cluster
- Sum of the squared sizes of representatives of c which are subgraph isomorphic to $G \rightarrow$ prefer high coverage
- For graphs that are no more supported by representatives $\rightarrow$ noise cluster
- Noise clusters do not need to have representatives.

CT-Index Fingerprint for Speeding Up

- CT-Index enumerates trees and circles up to a specific size and hashes the presence of these subgraphs into a binary fingerprint of fixed length.
- If G has set a bit to 1 but not H, then G cannot be subgraph isomorphic to H.
- We use this test as pre-filtering.

Cluster Splitting and Merging

- Cluster merging ensures a minimum separation
- Cluster separation is defined via the representative sets of two clusters
- $\rightarrow$ independent of cluster size
- Similarity of two representatives: size of their maximum common subgraph normalized by the size of the larger representative
- We merge two clusters, if they have many similar pairwise representatives

- Cluster splitting focuses on cluster homogeneity
- If the average coverage of the representatives of a cluster is too low, then we split the cluster
- coverage of a cluster is adaptive to the average coverage of all clusters
- all graphs in the cluster get part of the noise cluster
- after all splitting steps the noise cluster is pre-clustered

Empirical Evaluation

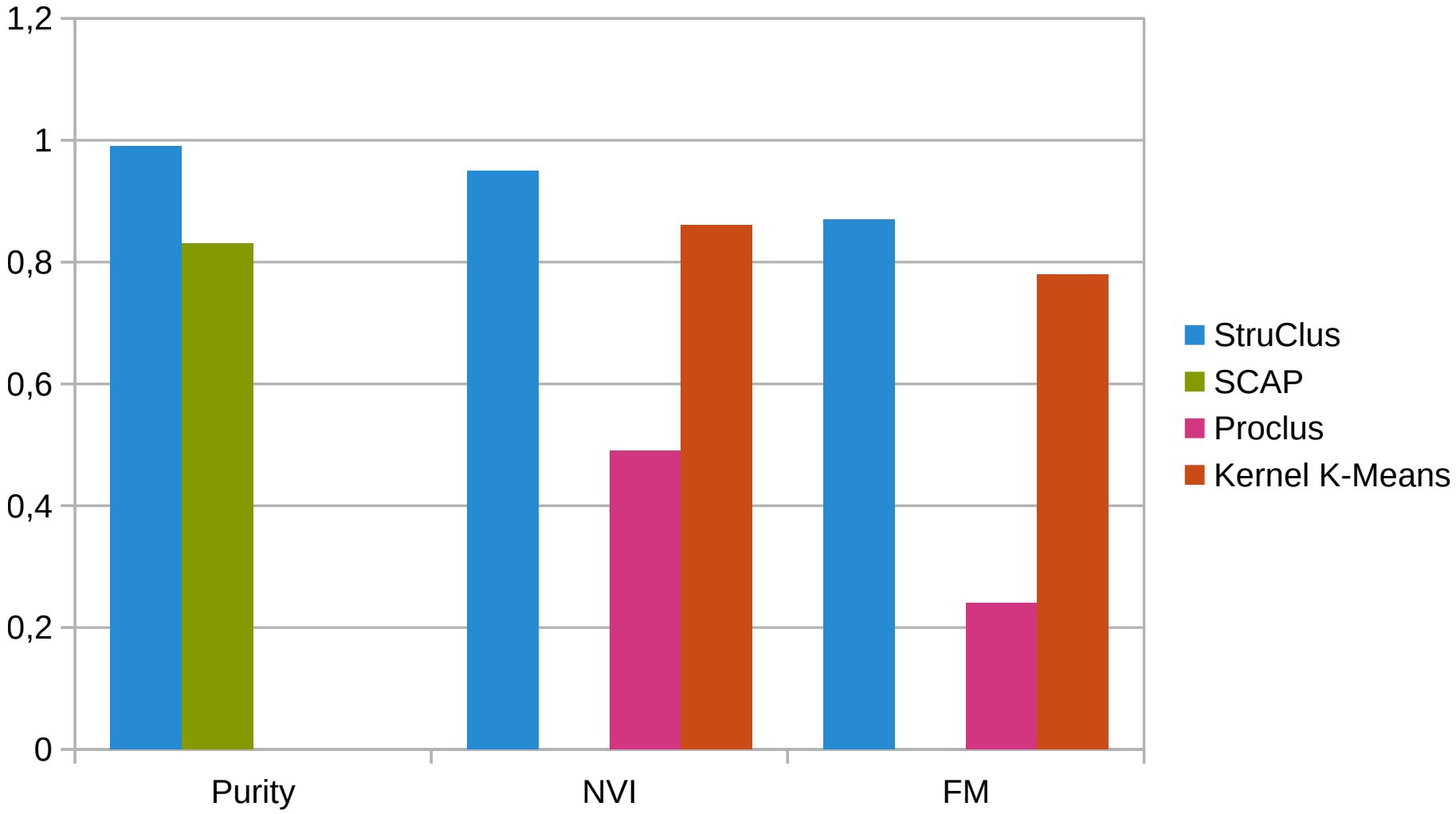
Synthetic Datasets

- 100 Clusters
- 5% noise graphs
- Combined seed patterns (cluster + common)
- Sizes: 10^3 to 10^6

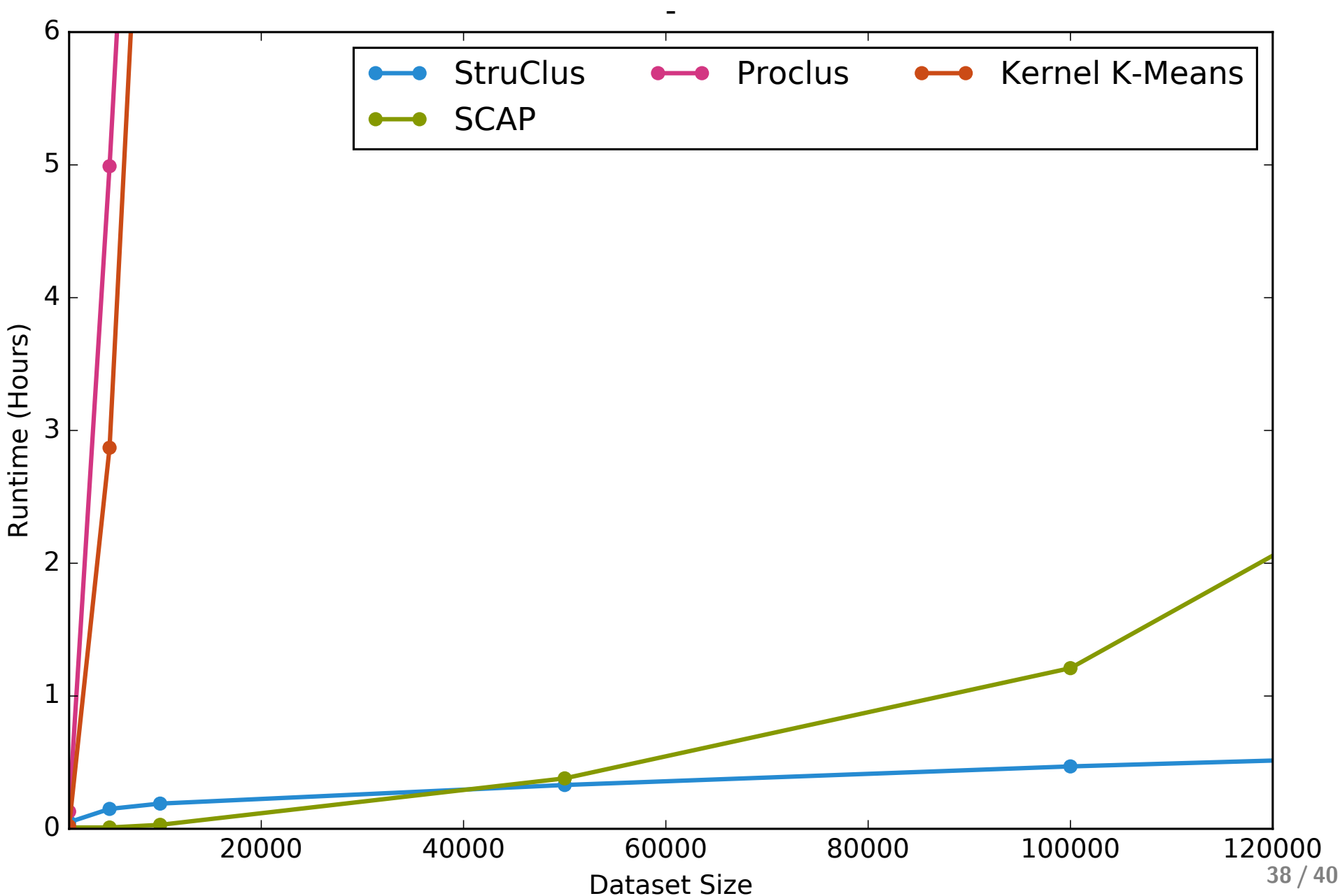
Competitors

- SCAP [Seeland, Karwath, Kramer, 2014]
- Proclus [Aggarwal, Procopiuc, Wolf, Yu, Park, 1999]
- Kernel K-Means [Girolami, 2002]

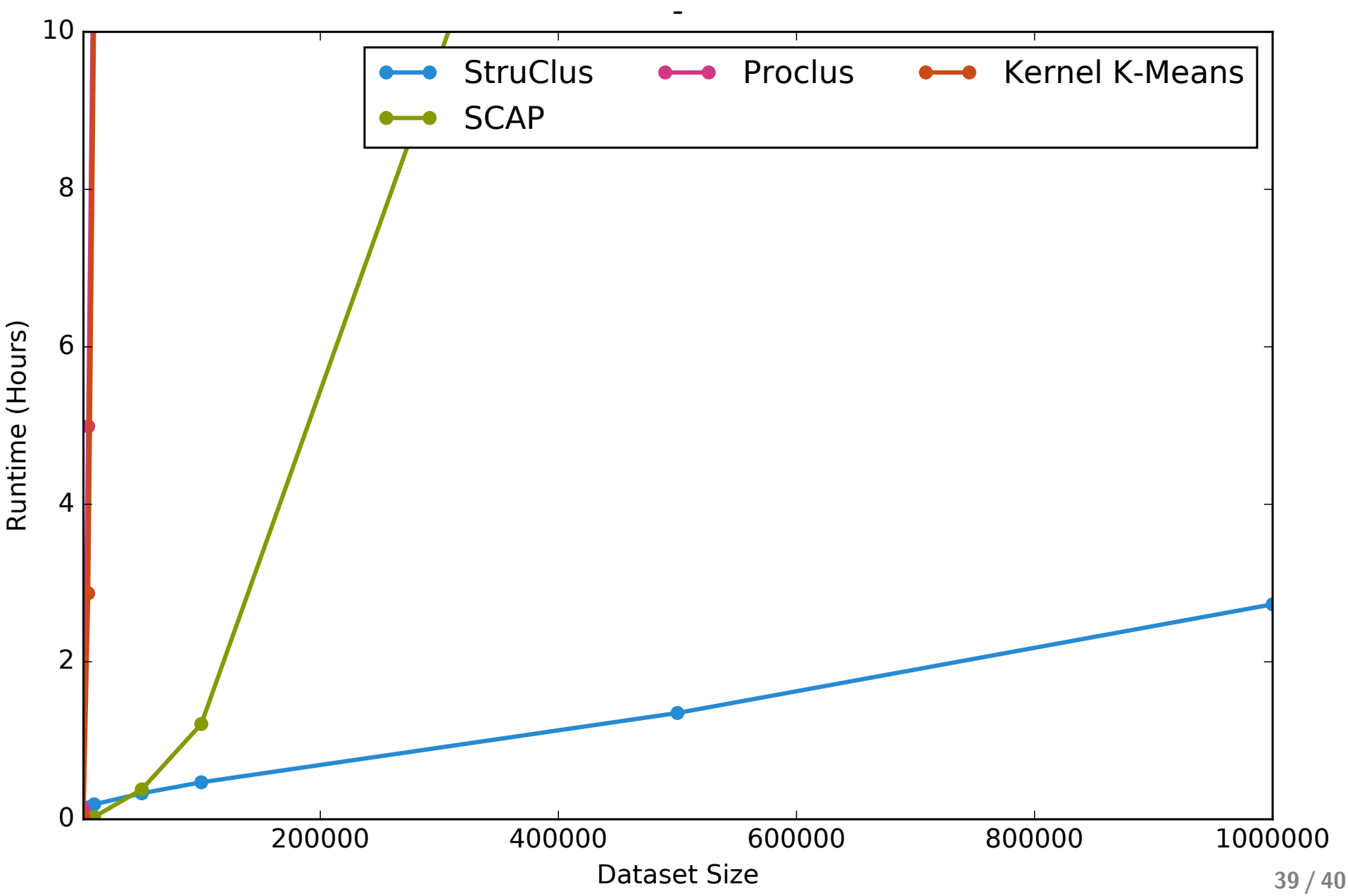
Quality


1,2
1
0,8
0,6
0,4
0,2
0
Purity
NVI
FM
StruClus
SCAP
Proclus
Kernel K-Means

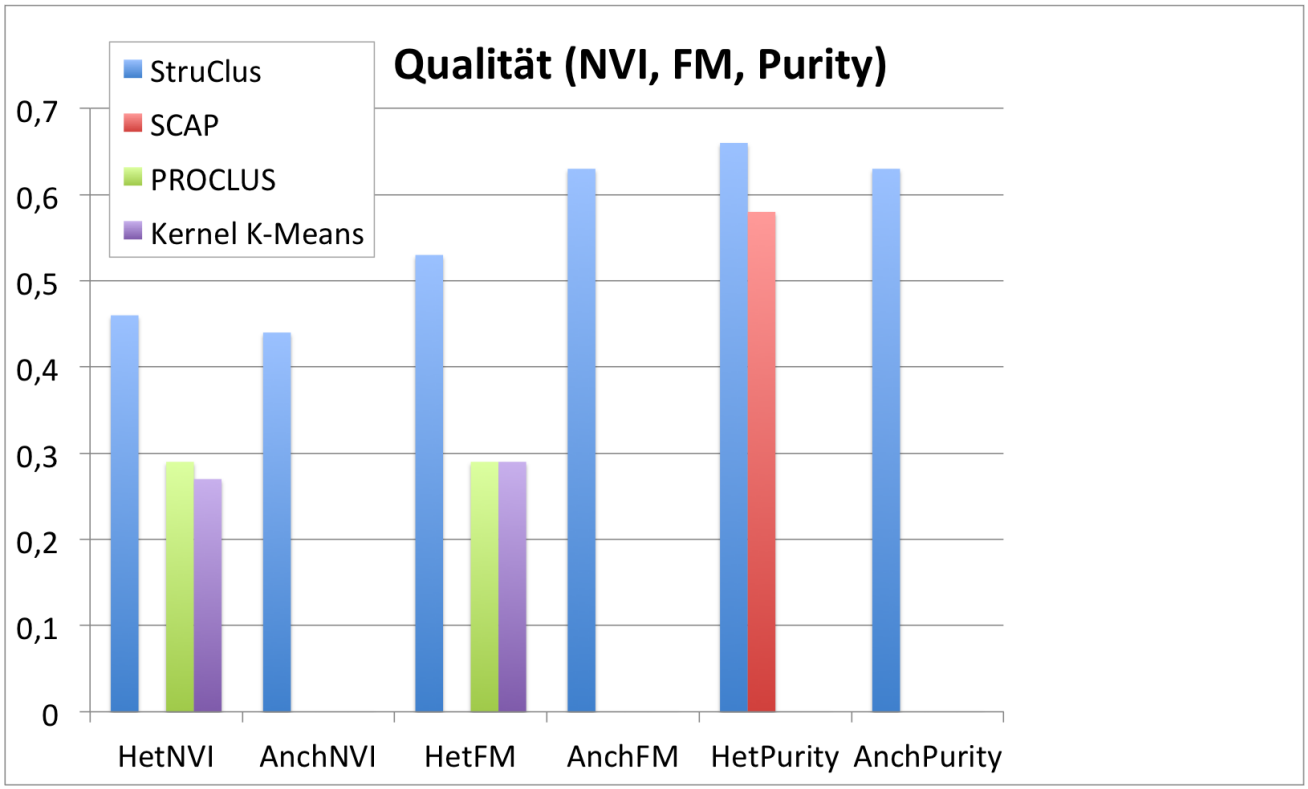
Runtime


StruClus
Proclus
Kernel K-Means
SCAP
Runtime (Hours)
0
1
2
3
4
5
6
20000
40000
60000
80000
100000
120000
Dataset Size

Runtime


StruClus
Proclus
Kernel K-Means
SCAP
Runtime (Hours)
Dataset Size
0
2
4
6
8
10
200000
400000
600000
800000
1000000

Real World Evaluation of the StruClus Algorithm



- Quality of the clustering for 3 molecule data bases:
- Heterocyclic: 10 000 (39), AnchorQuery: 65 700 (11)
- only StruClus: ChemDB (5 Mio. in 19 h)