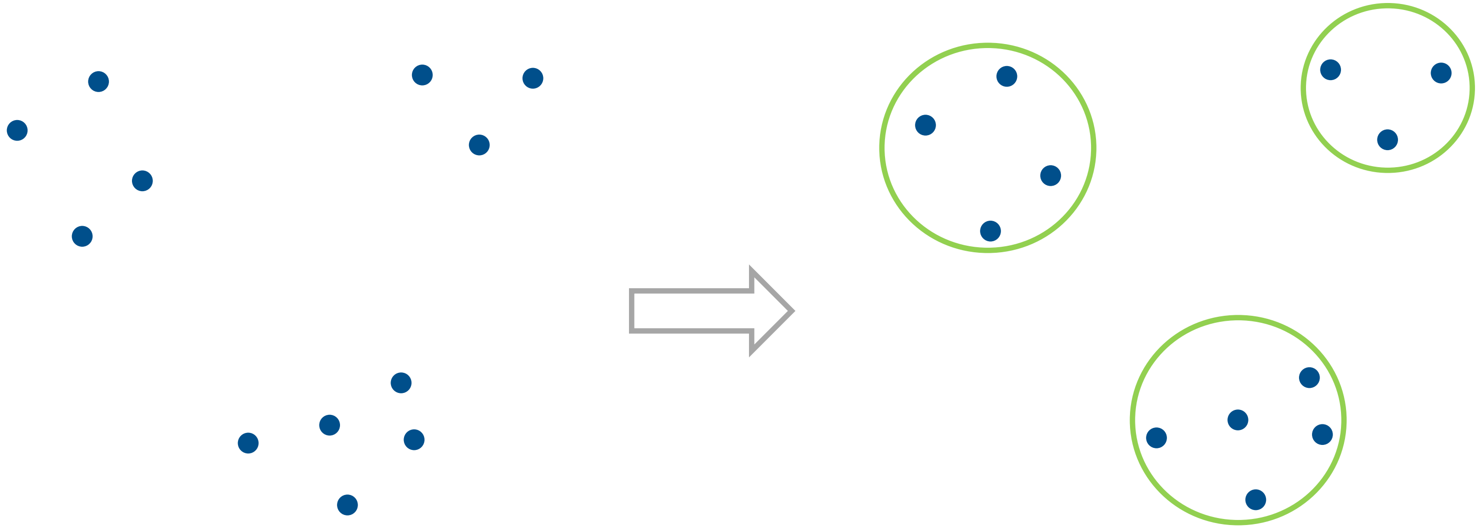
The Complexity of k-Means Clustering when Little is Known

Robert Ganian, Thekla Hamm,
Viktoriia Korchemna,
Karolina Okrasa, Kirill Simonov

Motivation

- Given the set of n points in d-dimensional space, group them into k clusters of "small size".
- One of the most impactful approaches in the area of data analysis and machine learning as a whole.



Means Clustering: Problem definition

Input:	**Matrix $\mathrm{A} \in D^{n \times d}$ over a domain $D \subseteq R$, integers k and l.**
Task:	Determine if there exists a matrix B over D containing at most k distinct rows such that $\|\|B - A\|\|^2 \leq l$

- We consider the Frobenius norm:

$$\|\|\mathrm{A}\|\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} A[i,j]^2$$

Means Clustering with Missing Entries (MCME)

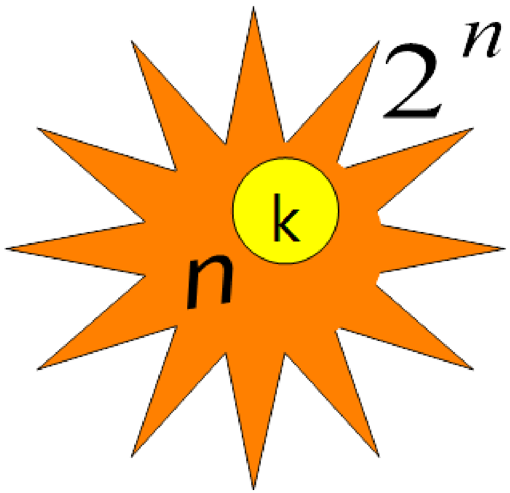
Input:	**Matrix $\mathrm{A} \in D^{n\times d}$ over a domain $D \subseteq R$, binary matrix W (the mask), integers k and l.**
Task:	Determine if there exists a matrix B over D containing at most k distinct rows such that $\lVert \mathrm{W} \circ (B - A) \rVert^2 \leq l$.

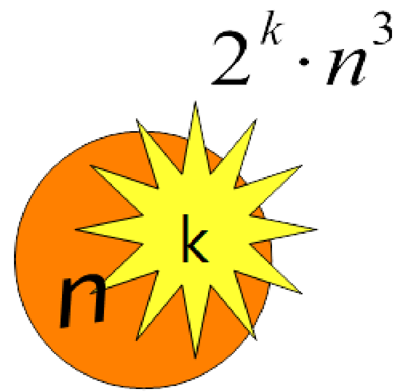
BOUNDED-DOMAIN MCME is NP-complete (Drineas et al., 2004; Aloise et al., 2009).

Parameterized complexity –> more fine-grained look into the complexity of the problem

Fixed-Parameter Tractability

– Can we identify **structural properties** of input which suffice for **tractability**?

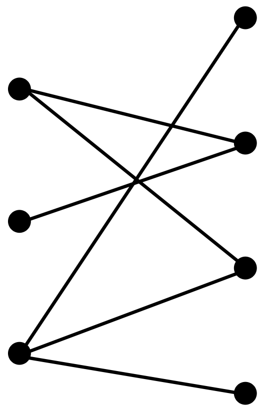




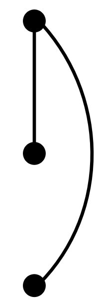
Definition: A problem is *fixed-parameter tractable (FPT)* when parameterized by an integer k (called the *parameter*) if it admits an algorithm with running time $f(k) \cdot n^{O(1)}$, where n is the size of the input and f is some computable function.

Input as a Graph

We use two graph representations of the mask W:



$$W = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$



Incidence graph G_I

$V_I = R_W \cup C_W$

$(i,j) \in E_I \Leftrightarrow W[i,j] = 1$

Primal graph G_P

$V_P = R_W$

$(i,j) \in E_P \Leftrightarrow (W[i], W[j]) > 0$

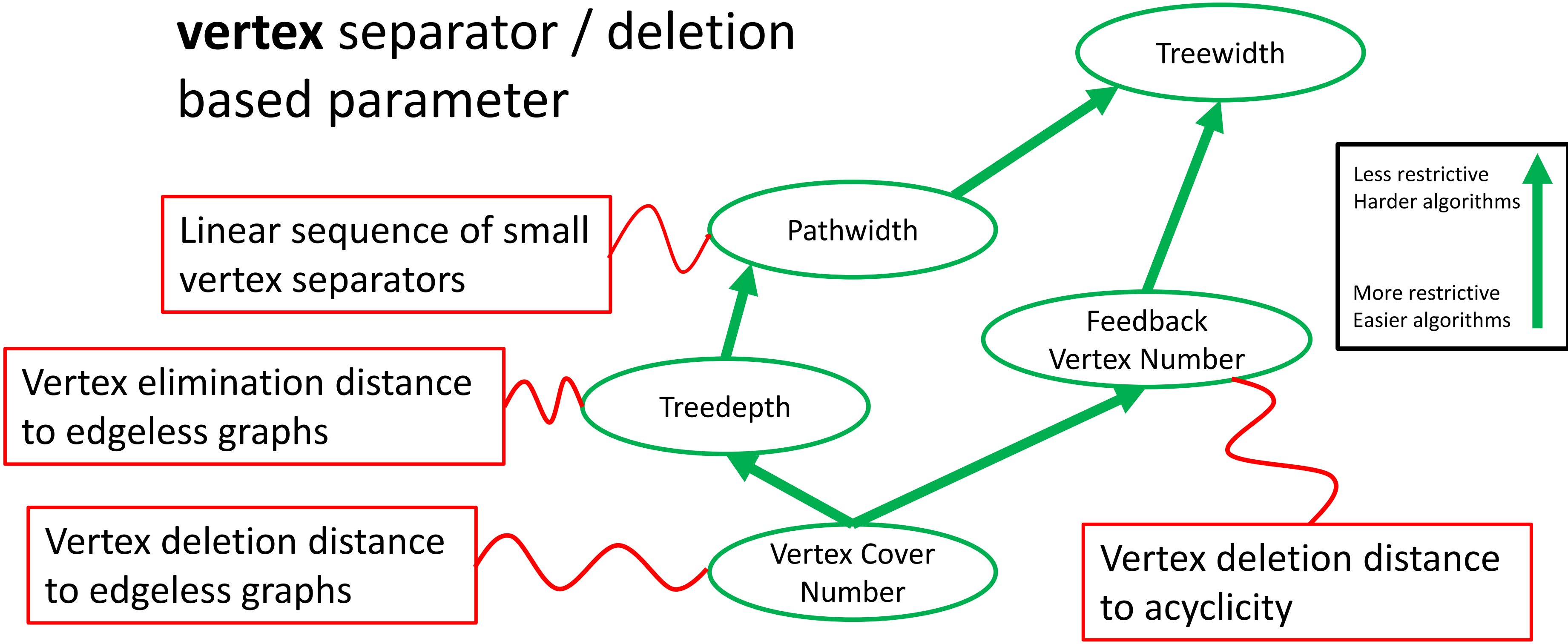
When Little is Known

- Eiben et al. (2021) and Ganian et al. (2018) obtained several fixed-parameter clustering algorithms by using a parameter called the *covering number*, which is the minimum number of rows and columns needed to cover all the *unknown* entries.
- What if most of the data is unknown or irrelevant?
- Can try: number of rows and columns to cover all the *known* entries = vertex cover number of G_I.
- In fact, we can do even better: *treewidth* of G_I ☺

Treewidth

- Treewidth iteratively decomposes the input along small **vertex separators**
  - The most general (the lest restrictive)

    vertex separator / deletion based parameter



New FPT Results

- incidence treewidth = treewidth of G_I

  (bounded domain)

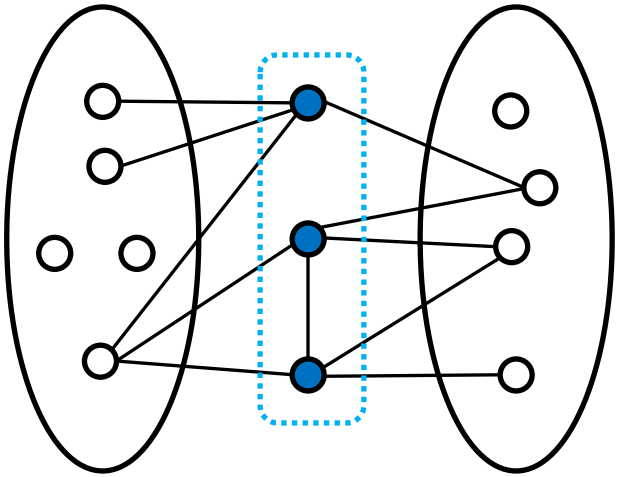
- primal treewidth = treewidth of G_P

  (more restrictive parameter, real-valued domain)

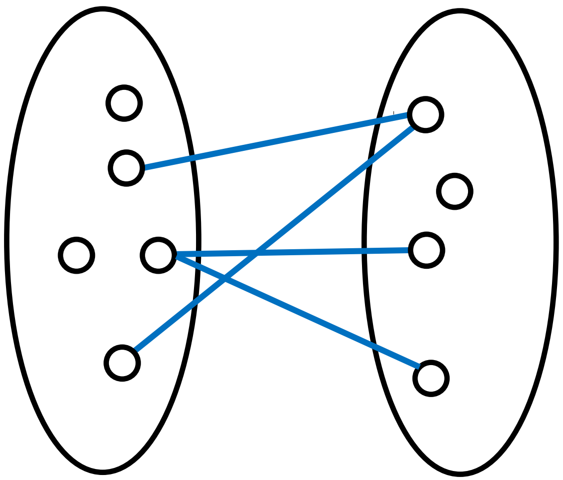
Is real-valued Means Clustering FPT when parameterized by d?
-> long-standing open problem dating back to Inaba et al.'s celebrated XP algorithm parameterized by k + d (1994).

Another approach: small edge cuts

– What about structural parameters that guarantee small **edge cuts**?



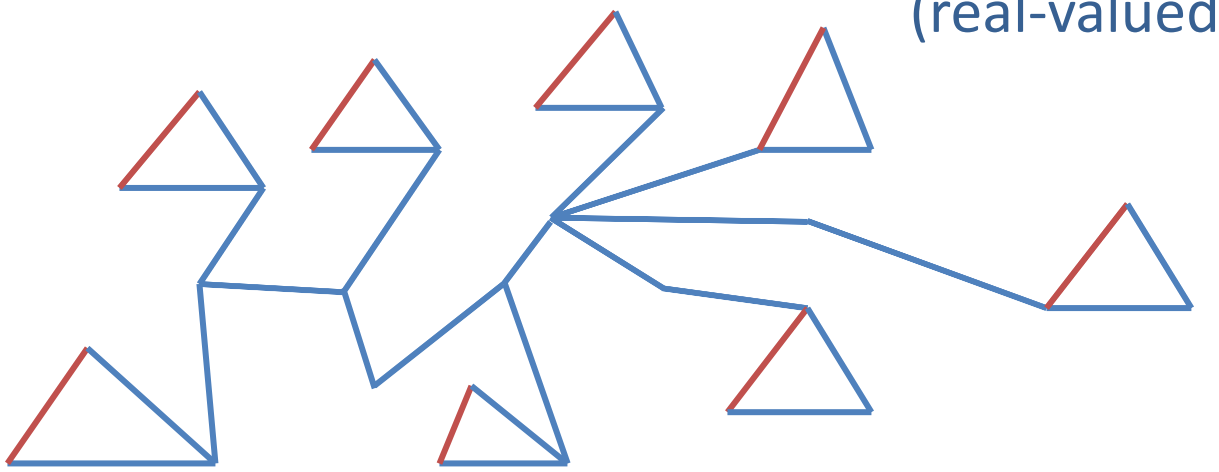
small vertex separator



small edge-cut

New FPT Results

- incidence treewidth = treewidth of G_I
  (bounded domain)
- primal treewidth = treewidth of G_P
  (real-valued domain)
- local feedback edge number of G_I
  (real-valued domain)



Thank you for your attention!