



## Cold-Start Active Correlation Clustering

Downloaded from: <https://research.chalmers.se>, 2026-05-12 18:10 UTC

Citation for the original published paper (version of record):

Aronsson, L., Wu, H., Haghiri Chehreghani, M. (2026). Cold-Start Active Correlation Clustering. Wsdm 2026 Proceedings of the 19th ACM International Conference on Web Search and Data Mining: 1068-1072. <http://dx.doi.org/10.1145/3773966.3779377>

N.B. When citing this work, cite the original published paper.



PDF Download  
3773966.3779377.pdf  
01 April 2026  
Total Citations: 0  
Total Downloads: 233

 Latest updates: <https://dl.acm.org/doi/10.1145/3773966.3779377>

SHORT-PAPER

## Cold-Start Active Correlation Clustering

**LINUS ARONSSON**, Chalmers University of Technology, Gothenburg, Vastra Gotaland, Sweden

**HAN WU**, Chalmers University of Technology, Gothenburg, Vastra Gotaland, Sweden

**MORTEZA HAGHIR CHEHREGHANI**, Chalmers University of Technology, Gothenburg, Vastra Gotaland, Sweden

Open Access Support provided by:

Chalmers University of Technology

Published: 21 February 2026

Citation in BibTeX format

WSDM '26: The Nineteenth ACM International Conference on Web Search and Data Mining  
February 22 - 26, 2026  
ID, Boise, USA

Conference Sponsors:

SIGKDD  
SIGWEB  
SIGIR  
SIGMOD

# Cold-Start Active Correlation Clustering

Linus Aronsson\*

Department of Computer Science and  
Engineering  
Chalmers University of Technology &  
University of Gothenburg  
Gothenburg, Sweden  
linaro@chalmers.se

Han Wu\*

Department of Computer Science and  
Engineering  
Chalmers University of Technology &  
University of Gothenburg  
Gothenburg, Sweden  
hanwu@student.chalmers.se

Morteza Haghir Chehreghani

Department of Computer Science and  
Engineering  
Chalmers University of Technology &  
University of Gothenburg  
Gothenburg, Sweden  
morteza.chehreghani@chalmers.se

## Abstract

We study active correlation clustering where pairwise similarities are not provided upfront and must be queried in a cost-efficient manner through active learning. Specifically, we focus on the cold-start scenario, where no true initial pairwise similarities are available for active learning. To address this challenge, we propose a coverage-aware method that encourages diversity early in the process. We demonstrate the effectiveness of our approach through several synthetic and real-world experiments.

## CCS Concepts

• **Computing methodologies** → **Active learning settings; Cluster analysis.**

## Keywords

Correlation clustering; active learning; cold-start learning.

### ACM Reference Format:

Linus Aronsson, Han Wu, and Morteza Haghir Chehreghani. 2026. Cold-Start Active Correlation Clustering. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining (WSDM '26)*, February 22–26, 2026, Boise, ID, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3773966.3779377>

## 1 Introduction

*Correlation clustering* (CC) [4, 10] clusters objects directly from the respective signed pairwise relations, accommodating both positive and negative similarities. CC has been used in diverse applications, including image segmentation [23], bioinformatics [7], spam filtering [5], social network analysis [1, 6, 34], duplicate detection [19], co-reference resolution [28], entity resolution [12], color naming [35], clustering aggregation [13, 18] and hierarchical clustering [9]. Computing the optimum is NP-hard and APX-hard [4, 10]; consequently, approximation strategies are employed in practice, with local-search variants often offering a favorable balance of quality and efficiency [16, 35]. In many real-world scenarios, the  $\binom{N}{2}$  pairwise similarities needed by CC are *not* available upfront. Obtaining them, e.g., from experts, crowd workers, or laboratory experiments, can be expensive and time-consuming [8, 11].

\*These authors contributed equally to this work.



This work is licensed under a Creative Commons Attribution 4.0 International License. *WSDM '26, Boise, ID, USA*

© 2026 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-2292-9/2026/02  
<https://doi.org/10.1145/3773966.3779377>

This motivates *active correlation clustering* (active CC), where the aim is to recover a high-quality CC solution while querying only a small fraction of pairs. We adopt the standard setting considered in prior work [3, 8, 11, 14, 25, 27, 32, 36]: (i) the objective is CC; (ii) pairwise similarities are unknown a priori; (iii) the algorithm may query a single (noisy) oracle under a fixed budget  $W \ll \binom{N}{2}$ ; and (iv) feature vectors are not assumed—information about the clustering is obtained solely from queried pairwise relations.

Early research proposed pivot-based algorithms with query-complexity guarantees under noise [27], adaptive variants of Kwik-Cluster [8, 11], and bandit-based formulations [14, 25]. While theoretically appealing, these approaches either rely on strong assumptions (e.g., known noise rates) or struggle in realistic noisy regimes. A flexible framework that decouples the query strategy from the downstream CC algorithm was later introduced in [3], enabling the design of general query strategies and the use of efficient local-search algorithms [16, 35]. Building on this framework, recent work introduced *information-theoretic query strategies* [2] (based on entropy and information gain) tailored to pairwise querying in CC and reported strong empirical improvements over existing baselines.

Despite their strengths, uncertainty-based methods, including information-theoretic approaches, have two key limitations. (i) In the *cold-start* setting, when no or few pairwise similarities are initially available, uncertainty estimates are unreliable, which can induce early *selection bias*: the algorithm repeatedly queries locally informative pairs in a small region before exploring the graph more broadly, delaying recovery of global structure. (ii) Under batch selection, they often pick highly redundant pairs within the same batch, a common issue in batch active learning [21, 29–31].

We address these challenges by proposing a *coverage-aware* query strategy for active CC that explicitly encourages diversity among queried pairs. Intuitively, the method prioritizes broad coverage by querying pairs that span many distinct objects. Our contributions are the following.

- We identify and empirically characterize the *cold-start sensitivity* of uncertainty-based query strategies in active CC, linking early-round failures to selection bias and insufficient coverage.
- We propose a simple and efficient coverage-aware method that prioritizes diversity in queried pairs. It offers two advantages: (i) it increases within-batch diversity, mitigating batch redundancy in batch active learning [29]; and (ii) it increases diversity across rounds, reducing selection bias and accelerating the acquisition of globally useful information.
- We demonstrate effectiveness and robustness on synthetic and real datasets, showing consistent gains in the cold-start setting.

**Algorithm 1** Generic Active CC

---

**Require:** Initial weights  $S^0$ , batch size  $B$ , total query budget  $W$ , query strategy  $\mathcal{S}$

- 1:  $i \leftarrow 0$ ;  $q \leftarrow 0$
- 2: **while**  $q < W$  **do**
- 3:    $\mathbf{c}^i \leftarrow \text{CC-ALGORITHM}(S^i)$
- 4:   Select a batch  $\mathcal{B} = \mathcal{S}(S^i, \mathbf{c}^i) \subseteq \mathcal{E}$  of size  $B$
- 5:   Query the oracle for all  $(u, v) \in \mathcal{B}$  and update the corresponding weights in  $S^{i+1}$
- 6:    $q \leftarrow q + |\mathcal{B}|$ ;  $i \leftarrow i + 1$
- 7: **end while**
- 8: **return**  $\mathbf{c}^i$

---

## 2 Active Correlation Clustering

In this section, we formalize active correlation clustering.

### 2.1 Problem Setup

Let  $\mathcal{V} = \{1, \dots, N\}$  be the set of vertices (objects) and  $\mathcal{E} = \{(u, v) \mid u, v \in \mathcal{V}, u < v\}$  the set of (undirected) edges. We consider a signed, weighted graph  $G = (\mathcal{V}, \mathcal{E}, S)$ , where  $S \in \mathbb{R}^{N \times N}$  is symmetric with zeros on the diagonal and entries  $S_{uv} \in [-1, 1]$  serving as *edge weights*: +1 indicates strong similarity, -1 strong dissimilarity, and values near 0 indicate uncertainty (including oracle ambiguity). Conceptually, CC operates on the complete signed graph; in the active setting only a small subset of weights is revealed by querying an oracle. We maintain an estimate  $S$  of the unknown ground-truth matrix  $S^*$ , updating entries as queries are answered.

A clustering is a partition of  $\mathcal{V}$ . We encode a clustering with  $K$  clusters as  $\mathbf{c} \in [K]^N$ , where  $c_u$  is the label of object  $u$ . We say a pair  $(u, v)$  *violates* a clustering  $\mathbf{c}$  if  $c_u = c_v$  and  $S_{uv} < 0$  or  $c_u \neq c_v$  and  $S_{uv} \geq 0$ . The CC objective penalizes violations and can be defined as  $R^{\text{CC}}(\mathbf{c} \mid S) = \sum_{(u,v) \in \mathcal{E}} |S_{uv}| \mathbb{I}[(u, v) \text{ violates } \mathbf{c}]$ . This is equivalent, up to an additive constant independent of  $\mathbf{c}$ , to the *max-correlation* form [15, 16]:  $R^{\text{MC}}(\mathbf{c} \mid S) = -\sum_{(u,v) \in \mathcal{E}} c_u = c_v S_{uv}$ . We have  $\arg\min_{\mathbf{c}} R^{\text{CC}}(\mathbf{c} \mid S) = \arg\min_{\mathbf{c}} R^{\text{MC}}(\mathbf{c} \mid S)$ . We therefore optimize  $R^{\text{MC}}$  (as it leads to a number of simplifications in the derived algorithms). The ground-truth clustering is  $\mathbf{c}^* = \arg\min_{\mathbf{c}} R^{\text{MC}}(\mathbf{c} \mid S^*)$ .

### 2.2 Active CC Procedure

We use the active CC procedure from [3] (Alg. 1): run a CC algorithm on the current signed graph, select a batch of pairs using query strategy  $\mathcal{S}$ , query the oracle, and update  $S$  until budget  $W$  is exhausted. We use the local-search CC algorithm from [3], due to its strong empirical performance. In the cold-start setting,  $S^0$  is uninformative (e.g., all zeros).

### 2.3 Information-Theoretic Methods

We here briefly recap the information-theoretic query strategies used in active CC, introduced by [2]. Let  $\mathcal{C}$  denote the set of all partitions of  $\mathcal{V}$ . We define the Gibbs distribution over clusterings with concentration  $\beta > 0$  as  $p^{\text{Gibbs}}(\mathbf{y} = \mathbf{c}) = \exp(-\beta R^{\text{MC}}(\mathbf{c} \mid S)) / Z$ , where  $Z = \sum_{\mathbf{c}' \in \mathcal{C}} \exp(-\beta R^{\text{MC}}(\mathbf{c}' \mid S))$  and  $\mathbf{y} \in \mathcal{C}$  is a random vector with sample space  $\mathcal{C}$ . Direct computation is intractable due to the enumeration of all clustering solutions in  $Z$ . We approximate

$p^{\text{Gibbs}}$  with a factorial distribution  $Q(\mathbf{y}) = \prod_{u \in \mathcal{V}} Q(y_u)$ , represented by  $\mathbf{Q} \in [0, 1]^{N \times K}$  with  $Q_{uk} = Q(y_u = k)$ . Using variational mean-field [17, 20], we alternate the synchronous updates  $\mathbf{Q} = \text{softmax}(-\beta \mathbf{M})$ , and  $\mathbf{M} = -\mathbf{S} \mathbf{Q}$  until convergence, where  $\mathbf{M} \in \mathbb{R}^{N \times K}$  is a matrix of assignment costs (i.e., element  $M_{uk}$  should be interpreted as the cost of assigning object  $u$  to cluster  $k$ ). The matrix  $\mathbf{M}$  can be initialized randomly. In short, this procedure converges to a local minimum of the KL-divergence between  $\mathbf{Q}$  and  $p^{\text{Gibbs}}$ . We refer to [2] for a detailed description.

**Entropy acquisition function.** Let  $E_{uv} \in \{0, 1\}$  be a random variable that indicates whether  $u$  and  $v$  are in the same cluster or not. The same-cluster probability is  $P(E_{uv} = 1) \approx \sum_{k=1}^K Q_{uk} Q_{vk}$ . The *entropy* acquisition function is defined as the entropy of  $E_{uv}$  [2]:

$$a^{\text{Entropy}}(u, v) := H(E_{uv}) = \mathbb{E}_{P(E_{uv})}[-\log P(E_{uv})]. \quad (1)$$

In this paper, we compare against  $a^{\text{Entropy}}$  to illustrate the issue of selection bias in uncertainty-based query strategies. We do not include acquisition functions based on expected information gain proposed by [2], for three main reasons: (i) they are also subject to selection bias, often more severely than entropy, (ii) their empirical performance is typically similar to entropy, and (iii) they are generally more computationally demanding in practice.

## 3 Coverage-Based Query Strategy

To mitigate cold-start selection bias and batch redundancy, we group edges into *query regions* and allocate the batch budget  $B$  across regions in proportion to their size-normalized informativeness. Regions can be *soft* (from the mean-field matrix  $\mathbf{Q}$ ) or *hard* (from the current clustering  $\mathbf{c}^i$ ). We unify both cases using a membership matrix  $\mathbf{U} \in [0, 1]^{N \times K}$ , where the hard case is given by  $U_{uk} = \mathbb{I}[c_u^i = k]$  for  $u \in \mathcal{V}$  and  $k \in [K]$ .

**Definition of query regions.** The set of query regions is a partition of the pairs  $\mathcal{E}$ . While the regions could be defined in many different ways, we propose to construct them given the current clustering solution  $\mathbf{c}^i \in \mathcal{C}$  with  $K$  clusters. We use  $\mathcal{R} = \{(a, a)\}_{a=1}^K \cup \{(a, b)\}_{1 \leq a < b \leq K}$  to represent the query regions. We then use  $R_{(a,a)} = \{(u, v) : c_u^i = c_v^i = a\}$  and  $R_{(a,b)} = \{(u, v) : \{c_u^i, c_v^i\} = \{a, b\}\}$  for  $a < b$  to denote the pairs in each region. This means that each region is either all pairs inside a cluster  $a \in [K]$ , or all pairs going between any two clusters (when  $a < b$ ). Notably, the number of clusters  $K$  can vary between iterations, since the CC algorithm used dynamically determines the number of clusters given the similarities queried so far. The regions in  $\mathcal{R}$  is thus adaptive to the iteration  $i$  of Alg. 1 both in terms of (i) which objects belong to each cluster, and (ii) the total number of clusters  $K$ .

**Query region sizes.** For any edge  $(u, v)$  and cluster indices  $a, b \in \{1, \dots, K\}$ , we define the region membership weights

$$w_{uv}^{(a,a)} = U_{ua} U_{va}, \quad w_{uv}^{(a,b)} = U_{ua} U_{vb} + U_{ub} U_{va} \text{ for } a < b. \quad (2)$$

Let  $\mathbf{s} = \mathbf{U}^T \mathbf{1}_N \in \mathbb{R}^K$  (each element is then  $s_a = \sum_u U_{ua}$ ) and  $\mathbf{B} = \mathbf{U}^T \mathbf{U}$ . The (soft) number of edges attributable to each region is  $N_{aa} = \sum_{u < v} w_{uv}^{(a,a)} = \frac{1}{2}(s_a^2 - B_{aa})$  and  $N_{ab} = \sum_{u < v} w_{uv}^{(a,b)} = s_a s_b - B_{ab}$  for  $(a < b)$ . If  $\mathbf{U}$  represents a hard assignment, i.e.,  $U_{ua} = \mathbb{I}\{c_u^i = a\}$ , then  $N_{aa} = |R_{(a,a)}|$  and  $N_{ab} = |R_{(a,b)}|$  for  $(a < b)$ . Thus, the region sizes reduce to the usual counts of within- and between-cluster pairs.

**Region informativeness mass.** Let  $\mathbf{A} \in \mathbb{R}_{\geq 0}^{N \times N}$  be a symmetric

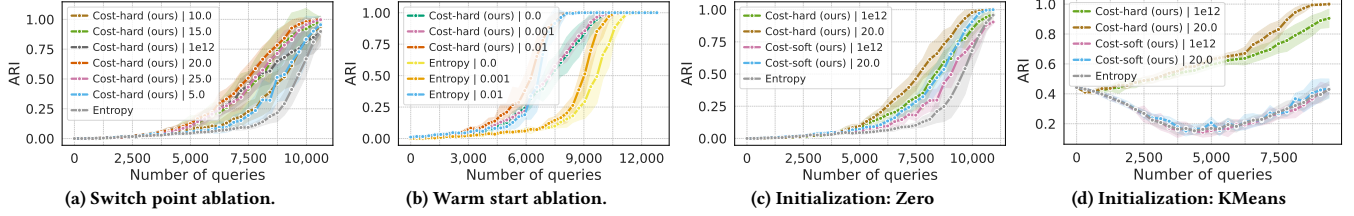


Figure 1: Ablation studies on the synthetic dataset. See Sec. 4 for a detailed description.

matrix, with  $A_{uu} = 0$ , where each element  $A_{uv}$  represents some notion of informativeness of the pair  $(u, v)$ . The total (soft) *informativeness mass* in each region is  $M_{aa} = \sum_{u < v} w_{uv}^{(a,a)} A_{uv} = \frac{1}{2} G_{aa}$  and  $M_{ab} = \sum_{u < v} w_{uv}^{(a,b)} A_{uv} = G_{ab}$  for  $a < b$  where  $\mathbf{G} = \mathbf{U}^\top \mathbf{A} \mathbf{U} \in \mathbb{R}^{K \times K}$ . We use the vectorized forms via  $\mathbf{G}$  in practice for efficiency. The purpose of defining a per-region value mass using an arbitrary matrix  $\mathbf{A}$  is to establish a flexible framework in which queries can be distributed across regions in any manner, thereby enabling a fully general and adaptable setup.

**Region informativeness normalized by region size.** We normalize by region size to avoid bias toward large regions to obtain the final score  $V_r = M_r / \max(N_r, \epsilon)$  for each region  $r \in \mathcal{R}$  ( $\epsilon > 0$  is used for stability). Then, the proportion of queries  $\pi_r \in [0, 1]$  (with  $\sum_r \pi_r = 1$ ) to be made in region  $r \in \mathcal{R}$  is computed as in Eq. (3).

$$\pi_r = \frac{V_r}{\sum_{s \in \mathcal{R}} V_s}. \quad (3)$$

**Choice of matrix  $\mathbf{A}$ .** We instantiate  $\mathbf{A}$  in several ways, depending on what we want the region proportions  $\{\pi_r\}$  to emphasize. (i) *Entropy*:  $A_{uv}^{\text{Entropy}} = a^{\text{Entropy}}(u, v)$  from Eq. (1), which will prioritize regions with large uncertainty according to the mean-field approximation  $\mathbf{Q}$ . (ii) *CC-cost contribution*:  $A_{uv}^{\text{Cost}} = |S_{uv}| \cdot \mathbb{I}[(u, v) \text{ violates } c^i]$  (based on the CC cost  $R^{\text{CC}}(c | S)$ ). This targets edges that are immediately relevant to reducing the CC objective. For example, if a cluster contains many negative edges (i.e., a high CC cost within the cluster), this likely indicates that the cluster should be split into two or more smaller clusters. Such inconsistencies can be resolved by querying additional similarities within the cluster. (iii) *Frequency*:  $A_{uv}^{\text{Freq}} = 1 - F_{uv}$  with  $F_{uv} \in \{0, 1\}$  indicating whether  $(u, v)$  has already been queried. This encourages broad coverage by prioritizing regions with many unqueried pairs relative to the region size. (iv) *Magnitude uncertainty (MU)*:  $A_{uv}^{\text{MU}} = 1 - |S_{uv}|$  (recall  $S_{uv} \in [-1, 1]$ ), giving higher scores to pairs whose similarity estimates are near 0. **Batch allocation and within-region selection.** Given region proportions  $\pi_r$  and batch size  $B$ , we allocate  $B_r = \text{round}(\pi_r B)$  queries to each region  $r$  and apply a largest-remainder adjustment to ensure  $\sum_r B_r = B$  and  $B_r \geq 0$ . For each region (e.g.,  $(a, b) \in \mathcal{R}$ ), we select  $B_r$  pairs from  $R_{(r)}$ ; if  $|R_{(r)}| < B_r$ , we query all pairs in

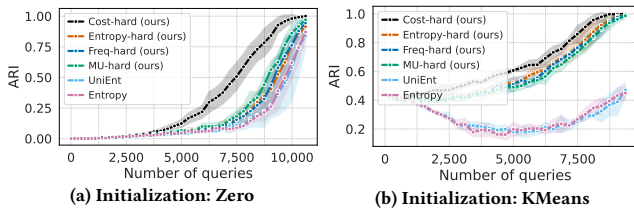


Figure 2: Comparison of diverse methods on synthetic dataset.

$R_{(r)}$  and reassign the leftover budget to other regions. Within region  $r$ , pairs are sampled without replacement with  $p(u, v | r) = a^{\text{Entropy}}(u, v) / \sum_{(w, z) \in R_{(r)}} a^{\text{Entropy}}(w, z)$ . Equivalently, this amounts to selecting, among pairs  $(u, v) \in R_{(r)}$ , the top- $B_r$  pairs according to the score  $a(u, v) = \log(a^{\text{Entropy}}(u, v)) + \epsilon_{uv}$ , where  $\epsilon_{uv} \sim \text{Gumbel}(0, 1)$ . This combines uncertainty-driven selection with exploration and empirically outperforms taking the top- $B_r$  pairs by  $a^{\text{Entropy}}$  directly. Combining the methods for computing region proportions (soft or hard) with a given matrix  $\mathbf{A}$  (Entropy, Cost, Freq, or MU) yields 8 variants.

## 4 Experiments

We follow the experimental protocol of [2]. We evaluate on one synthetic dataset (10 balanced clusters) and five real-world datasets: CIFAR-10 [24], 20 Newsgroups [22], Forest Type Mapping [22], User Knowledge Modeling [22], and MNIST [26]. Unless stated otherwise, experiments use the synthetic dataset and at most  $N = 1000$  instances because some baselines are expensive (our methods scale to larger datasets). Preprocessing follows [2], except that 20 Newsgroups uses samples from all 20 topics.

In addition, we follow [2] and adopt the same CC algorithm, noisy oracle, evaluation metric, and baselines. The oracle returns the ground-truth similarity (+1 if two instances belong to the same class and  $-1$  otherwise) with probability  $1 - \gamma$ , and a random value in  $[-1, +1]$  with probability  $\gamma$ , where we fix  $\gamma = 0.4$ . At each iteration of the active CC procedure, we compute the adjusted rand index (ARI) between  $c^i$  and the ground-truth clustering (given by the true class labels of each dataset). The baselines include entropy from [2] (Eq. (1)), where we apply the sampling approach described at the end of Sec. 3 to improve batch diversity, following [2]; maxmin and maxexp from [3], which originally introduced the active CC procedure in Alg. 1; a pivot-based active CC algorithm called QECC [11]; two adapted state-of-the-art active constraint clustering methods COBRAS [36] and nCOBRAS [33]; and a recent bandit-based approach KC-FB [25]. Finally, we include a simple baseline, denoted *UniEnt*, that selects pairs randomly for a few iterations before switching to entropy. After empirical tuning on each dataset, we fix the number of iterations before switching to 20 for the synthetic dataset and 10 for the real-world datasets. This baseline highlights that our approach outperforms naive random exploration, a common strategy for mitigating selection bias. Importantly, we query each pair at most once.

We consider two strategies for initializing the similarity matrix  $S^0$ : (i) all similarities are set to zero, representing no prior knowledge; and (ii) we apply  $k$ -means clustering on the feature vectors of each dataset and set  $S_{uv}^0 = 0.01$  if  $(u, v)$  are assigned to the same cluster and  $-0.01$  otherwise. The second approach incorporates weak prior

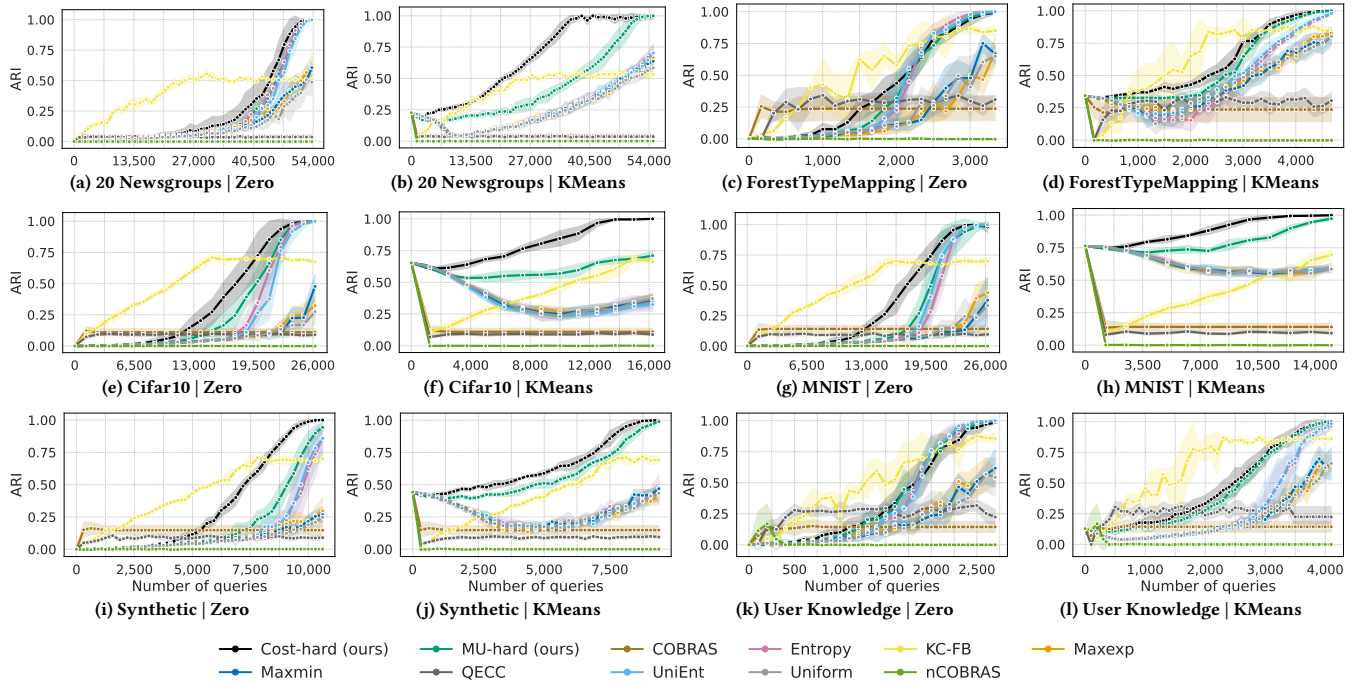


Figure 3: Results for different methods across datasets.

knowledge about the true clustering but may introduce bias if the feature space is noisy, potentially leading to selection bias. Unless otherwise specified, we use the zero initialization.

It is reasonable to assume that once sufficient information about the true similarities has been collected, one can safely *switch* to a purely uncertainty-driven strategy without suffering from selection bias. Our first experiment investigates this hypothesis (Fig. 1a) by evaluating the performance of our method *cost-hard* when switching to entropy at different iterations. For reference, we also include pure entropy (i.e., starting from iteration 0). We find that our method consistently outperforms pure entropy across all switch points, demonstrating robustness to the choice of when to switch. This highlights the potential for future work on dynamically determining the optimal switch point. Empirically, switching after 20 iterations yields the best performance, surpassing even the case of never switching (1e12). Based on these findings, we fix the switch point to 20 for the synthetic dataset and 10 for all real-world datasets in the remaining experiments (empirically chosen).

In the next experiment (Fig. 1b), we study *warm-start* by varying the proportion of ground-truth similarities revealed at initialization and comparing *cost-hard* to entropy. Entropy performs well with substantial initial information (0.01) but degrades sharply with limited warm-start (0 or 0.001) due to selection bias, whereas *cost-hard* remains more robust. Importantly, this ablation assumes access to perfect (noise-free) oracle information, which is unrealistic in practice and underscores the need for methods that perform well in the cold-start regime. Moreover, even a 0.01% warm-start at  $N = 5000$  corresponds to roughly 125,000 pairs known in advance.

In Figs. 1c–1d, we compare *soft* vs. *hard* region memberships under two switch points. Hard regions perform better for both initializations; with *k*-means initialization, the soft variant shows

selection bias similar to entropy, likely due to its reliance on uncertainty estimates from  $Q$ . We therefore use hard memberships in the remaining experiments. In Fig. 2, we compare choices of  $A$  (cost, entropy, freq, MU): *cost-hard* performs best overall, followed by MU-hard, and we focus on these two in remaining experiments. UniEnt is consistently outperformed by our methods, indicating stronger initial exploration than random querying.

Finally, Fig. 3 presents the results for all methods across all datasets and both initialization strategies. Overall, we observe that our methods reach  $ARI = 1$  more quickly than the baseline methods on most datasets, demonstrating the effectiveness of our approach in cold-start scenarios.

## 5 Conclusion

We proposed a coverage-aware query strategy for cold-start active correlation clustering that promotes diversity in the selected pairwise similarities. Experiments on synthetic and real datasets showed that our methods consistently reduce selection bias and discovers the ground-truth clustering faster than existing baselines.

## Acknowledgments

The work of Linus Aronsson and Morteza Haghir Chehreghani was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Finally, the computations and data handling was enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

## Ethical Considerations

This work focuses on methodological advances in active correlation clustering. While the research itself does not directly process sensitive personal data, potential applications in domains such as social networks, bioinformatics, or image data raise ethical considerations. Possible negative societal impacts include risks of reinforcing biases present in the data, privacy concerns when clustering sensitive information, and misuse of the technology for surveillance or discriminatory purposes. To mitigate such risks, practitioners should carefully consider the choice of datasets, ensure appropriate anonymization where personal data is involved, and evaluate fairness and robustness of clustering outcomes. As our method is intended as a general-purpose algorithmic contribution, we emphasize the responsibility of downstream users to apply it in ethically sound and socially beneficial contexts.

## References

- [1] Linus Aronsson and Morteza Haghir Chehreghani. 2025. An Efficient Local Search Approach for Polarized Community Discovery in Signed Networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- [2] Linus Aronsson and Morteza Haghir Chehreghani. 2025. Information-Theoretic Active Correlation Clustering. In *2025 IEEE International Conference on Data Mining (ICDM)*.
- [3] Linus Aronsson and Morteza Haghir Chehreghani. 2024. Correlation Clustering with Active Learning of Pairwise Similarities. *Transactions on Machine Learning Research* (2024).
- [4] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation Clustering. *Machine Learning* 56, 1-3 (2004), 89–113.
- [5] Francesco Bonchi, David García-Soriano, and Edo Liberty. 2014. Correlation clustering: from theory to practice. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- [6] Francesco Bonchi, Aristides Gionis, Francesco Gullo, and Antti Ukkonen. 2012. Chromatic Correlation Clustering. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1321–1329.
- [7] Francesco Bonchi, Aristides Gionis, and Antti Ukkonen. 2013. Overlapping correlation clustering. *Knowl. Inf. Syst.* 35, 1 (2013), 1–32.
- [8] Marco Bressan, Nicolò Cesa-Bianchi, Andrea Paudice, and Fabio Vitale. 2019. Correlation Clustering with Adaptive Similarity Queries. In *Advances in Neural Information Processing Systems*, Vol. 32.
- [9] Morteza Haghir Chehreghani and Mostafa Haghir Chehreghani. 2024. Hierarchical Correlation Clustering and Tree Preserving Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 23083–23093.
- [10] Erik D. Demaine, Dotan Emanuel, Amos Fiat, and Nicole Immorlica. 2006. Correlation clustering in general weighted graphs. *Theor. Comput. Sci.* 361, 2-3 (2006), 172–187.
- [11] David García-Soriano, Konstantin Kutzkov, Francesco Bonchi, and Charalampos Tsourakakis. 2020. Query-Efficient Correlation Clustering. In *Proceedings of The Web Conference*. 1468–1478.
- [12] Lise Getoor and Ashwin Machanavajjhala. 2012. Entity Resolution: Theory, Practice & Open Challenges. *Proc. VLDB Endow.* 5 (2012), 2018–2019.
- [13] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. 2007. Clustering aggregation. *ACM Trans. Knowl. Discov. Data* 1, 1 (2007), 4.
- [14] Francesco Gullo, Domenico Mandaglio, and Andrea Tagarelli. 2023. A combinatorial multi-armed bandit approach to correlation clustering. *Data Min. Knowl. Discov.* 37, 4 (2023), 1630–1691.
- [15] Morteza Haghir Chehreghani. 2013. *Information-theoretic validation of clustering algorithms*. Ph. D. Dissertation.
- [16] Morteza Haghir Chehreghani. 2023. Shift of pairwise similarities for data clustering. *Mach. Learn.* 112, 6 (2023), 2025–2051.
- [17] Morteza Haghir Chehreghani, Alberto Giovanni Busetto, and Joachim M. Buhmann. 2012. Information Theoretic Model Validation for Spectral Clustering. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, Vol. 22. 495–503.
- [18] Morteza Haghir Chehreghani and Mostafa Haghir Chehreghani. 2020. Learning representations from dendrograms. *Mach. Learn.* 109 (2020).
- [19] Otkie Hassanzadeh, Fei Chiang, Renée J. Miller, and Hyun Chul Lee. 2009. Framework for Evaluating Clustering Algorithms in Duplicate Detection. *Proc. VLDB Endow.* 2, 1 (2009), 1282–1293.
- [20] Thomas Hofmann, Jan Puzicha, and Joachim M. Buhmann. 1998. Unsupervised Texture Segmentation in a Deterministic Annealing Framework. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 8 (1998), 803–818.
- [21] Sanna Jarl, Linus Aronsson, Sadegh Rahrovani, and Morteza Haghir Chehreghani. 2022. Active learning of driving scenario Trajectories. *Eng. Appl. Artif. Intell.* 113 (2022), 104972.
- [22] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. 2023. The UCI Machine Learning Repository.
- [23] Sungwoong Kim, Sebastian Nowozin, Pushmeet Kohli, and Chang Dong Yoo. 2011. Higher-Order Correlation Clustering for Image Segmentation. In *Advances in Neural Information Processing Systems 24 (NIPS)*. 1530–1538.
- [24] Alex Krizhevsky. 2009. *Learning multiple layers of features from tiny images*. Technical Report.
- [25] Yuko Kuroki, Atsushi Miyauchi, Francesco Bonchi, and Wei Chen. 2024. Query-Efficient Correlation Clustering with Noisy Oracle. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- [26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- [27] Arya Mazumdar and Barna Saha. 2017. Clustering with Noisy Queries. In *Advances in Neural Information Processing Systems*, Vol. 30.
- [28] Andrew McCallum and Ben Wellner. 2004. Conditional Models of Identity Uncertainty with Application to Noun Coreference. In *Advances in Neural Information Processing Systems*. 905–912.
- [29] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A Survey of Deep Active Learning. *ACM Comput. Surv.* 54, 9 (2021).
- [30] Peter Samoaa, Linus Aronsson, Philipp Leitner, and Morteza Haghir Chehreghani. 2023. Batch Mode Deep Active Learning for Regression on Graph Data. In *2023 IEEE International Conference on Big Data (BigData)*. 5904–5913.
- [31] Peter Samoaa, Linus Aronsson, Antonio Longa, Philipp Leitner, and Morteza Haghir Chehreghani. 2024. A unified active learning framework for annotating graph data for regression tasks. *Engineering Applications of Artificial Intelligence* 138 (2024), 109383.
- [32] Sandeep Silwal, Sara Ahmadian, Andrew Nystrom, Andrew McCallum, Deepak Ramachandran, and Seyed Mehran Kazemi. 2023. KwikBucks: Correlation Clustering with Cheap-Weak and Expensive-Strong Signals. In *International Conference on Learning Representations*.
- [33] Jonas Soenen, Sebastijan Dumancic, Hendrik Blockeel, Toon Van Craenendonck, F Hutter, K Kersting, J Lijffijt, and I Valera. 2021. Tackling noise in active semi-supervised clustering. 121 - 136 pages.
- [34] Jiliang Tang, Yi Chang, Charu Aggarwal, and Huan Liu. 2016. A Survey of Signed Network Mining in Social Media. *ACM Comput. Surv.* 49, 3 (2016).
- [35] Erik Thiel, Morteza Haghir Chehreghani, and Devdatt P. Dubhashi. 2019. A Non-Convex Optimization Approach to Correlation Clustering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*. 5159–5166.
- [36] Toon van Craenendonck, Sebastijan Dumancic, Elia Van Wolputte, and Hendrik Blockeel. 2018. COBRAS: Interactive Clustering with Pairwise Queries. In *International Symposium on Intelligent Data Analysis*.