Relaxed Oracles for Semi-Supervised Clustering

Taewan Kim
The University of Texas at Austin
twankim@utexas.edu

Joydeep Ghosh
The University of Texas at Austin
jghosh@utexas.edu

Abstract

Pairwise "same-cluster" queries are one of the most widely used forms of supervision in semi-supervised clustering. However, it is impractical to ask human oracles to answer every query correctly. In this paper, we study the influence of allowing "not-sure" answers from a weak oracle and propose an effective algorithm to handle such uncertainties in query responses. Two realistic weak oracle models are considered where ambiguity in answering depends on the distance between two points. We show that a small query complexity is adequate for effective clustering with high probability by providing better pairs to the weak oracle. Experimental results on synthetic and real data show the effectiveness of our approach in overcoming supervision uncertainties and yielding high quality clusters. [1]

1 Introduction

Clustering is one of the most popular procedures for extracting meaningful insights from unlabeled data. However, clustering is also very challenging for a wide variety of reasons [14]. Finding the optimal solution of even the simple k-means objective is known to be NP-hard [13, 17, 23, 20]. Second, the quality of a clustering algorithm is difficult to evaluate without context. Semi-supervised clustering is one way to overcome these problems by providing a small amount of additional knowledge related to the task [9, 12, 10, 11, 6, 16, 18, 4, 19, 1].

The semi-supervised active clustering (SSAC) framework proposed by Ashtiani et al. [4] combines both margin property and pairwise constraints in the active query setting. A domain expert can help clustering by answering same-cluster queries, which ask whether two samples belong to the same cluster or not. By using an algorithm with two phases, it was shown that the oracle's clustering can be recovered in polynomial time with high probability. However, their formulation of the same-cluster query has only two choices of answers, *yes* or *no*. This might be impractical as a domain expert can also encounter ambiguous situations which are difficult to respond to in a short time.

Our work is motivated by the following question: "Is it possible to perform a clustering task efficiently even with a non-ideal domain expert?". We answer this question by formulating practical weak oracle models and allowing *not-sure* answers to query responses. Our model assumptions considers two reasonable scenarios that may lead to ambiguity in answering a same-cluster query: (i) distance between two points from different clusters is too small, and (ii) distance between two points within the same cluster is too large. We prove that our improved SSAC algorithm can work well under uncertainties if there exists at least one cluster element close enough to the center.

Experimental results on both synthetic and real data show the effective performance of our approach. In particular, our algorithm successfully deals with uncertainties compared to the previous SSAC algorithm by relaxing an oracle's role and providing better pairs for annotation in an active semi-supervision framework.

[1] This paper focuses on the distance-based weak oracle models with additional experimental results. Proofs for theoretical results are available in the extended version. [15]

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Problem Setting

For the purpose of theoretical analysis, the domain of data is assumed to be the Euclidean space $\mathbb{R}^m$, and each center of a clustering $\mathcal{C}$ is defined as a mean of elements in the corresponding cluster, i.e. $\mu_i = \frac{1}{|C_i|}\sum_{x\in C_i} x, \forall i \in [k]$. Then, an optimal solution of the k-means clustering is a center-based clustering.[2] Also, a γ-margin property ensures the existence of an optimal clustering.

Definition 1 (Center-based clustering)**.** *A clustering $\mathcal{C} = \{C_1, \cdots, C_k\}$ is a center-based clustering of $\mathcal{X} \subset \mathbb{R}^m$ with k clusters, if there exists a set of centers $\mu = \{\mu_1, \cdots, \mu_k\} \subset \mathbb{R}^m$ satisfying the following condition with a distance metric $d(x, y)$:*

$$x \in C_i \Leftrightarrow i = \arg\min_j d(x, \mu_j), \ \ \forall x \in \mathcal{X} \text{ and } i \in [k]$$

Definition 2 (γ-margin property - Clusterability)**.** *Let $\mathcal{C}$ be a center-based clustering of $\mathcal{X}$ with clusters $\mathcal{C} = \{C_1, \cdots, C_k\}$ and corresponding centers $\{\mu_1, \cdots, \mu_k\}$. $\mathcal{C}$ satisfies the γ-margin property if the following condition is true:*

$$\gamma d(x, \mu_i) < d(y, \mu_i), \ \ \forall i \in [k], \forall x \in C_i, \forall y \in \mathcal{X} \setminus C_i$$

Problem Formulation We apply the SSAC algorithm on data $\mathcal{X}$, which is supported by a weak oracle that receives weak same-cluster queries. The true clustering $\mathcal{C}$ satisfies the γ-margin property.

Definition 3 (Weak Same-cluster Query)**.** *A weak same-cluster query asks whether two data points $x_1, x_2 \in \mathcal{X}$ belong to the same cluster and receives one of three responses from an oracle.*

$$Q(x_1, x_2) = \begin{cases} 1 & \text{if } x_1, x_2 \text{ are in the same cluster} \\ 0 & \text{if not-sure} \\ -1 & \text{if } x_1, x_2 \text{ are in different clusters} \end{cases}$$

Definition 4 (Weak Pairwise Cluster-assignment Query)**.** *A weak pairwise cluster-assignment query identifies the cluster index of a given data point x by asking k weak same-cluster queries $Q(x, y_i)$, where $y_i \in C_{\pi(i)}$, $i \in [k]$. One of $k+1$ responses is inferred from an oracle with $\mathcal{C} = \{C_1, \cdots, C_k\}$. $\pi(\cdot)$ is a permutation defined on $[k]$ which is determined during the assignment process accordingly.*

$$Q(x) = \begin{cases} t & \text{if } x \in C_{\pi(t)}, t \in [k] \\ 0 & \text{if not-sure} \end{cases}$$

In our framework, the cluster-assignment process uses k weak same-cluster queries and therefore only depends on pairwise information provided by weak oracles. And we denote the radius of a cluster as $r(C_i) \triangleq \max_{x\in C_i} d(x, \mu_i)$ throughout the paper.

Algorithm 1 SSAC for Weak Oracles

Input: Dataset $\mathcal{X}$, an oracle for weak query Q, target number of clusters k, sampling numbers (η, β), and a parameter $\delta \in (0, 1)$.

$\mathcal{C} = \{\}, \ \mathcal{S}_1 = \mathcal{X}, \ r = \lceil k\eta \rceil$
for $i = 1$ to k **do**
  - Phase 1:
    $Z \sim \text{Uniform}(\mathcal{S}_i, r)$ // Draw r samples from S_i
    for $1 \le t \le k$ **do**
      $Z_t = \{x \in Z : Q(x) = t\}$ // Pairwise cluster-assignment query
    end for
    $p = \arg\max_t |Z_t|$, $\mu'_p \triangleq \frac{1}{|Z_p|}\sum_{x\in Z_p} x$
  - Phase 2:
    $\hat{\mathcal{S}}_i = \text{sorted}(\mathcal{S}_i)$ // Increasing order of $d(x, \mu'_p), \ x \in \mathcal{S}_i$
    $r'_i = \text{BinarySearch}(\hat{\mathcal{S}}_i, Z_p, \mu'_p, \beta)$ // Same-cluster query
    $C'_p = \{x \in \mathcal{S}_i : d(x, \mu'_p) < r'_i\}, \ \mathcal{S}_{i+1} = \mathcal{S}_i \setminus C'_p, \ \mathcal{C} = \mathcal{C} \cup \{C'_p\}$
end for

Output: A clustering $\mathcal{C}$ of the set $\mathcal{X}$

[2] In fact, this will hold for all Bregman divergences [8].

3 SSAC with Distance-Weak Oracles

It is reasonable to expect the accuracy of feedback from domain experts to depend on the inherent ambiguities of the given pairs of samples. The cause of "not-sure" answer for the same-cluster query can be investigated based on the distance between the elements in a feature space. Two reasons for having indefinite answers are considered in this work: (i) points from different clusters are too close, and (ii) points within the same cluster are too far. The first situation happens a lot in the real world. For instance, distinguishing wolves from dogs is not an easy task if a Siberian Husky is considered. The second case is also reasonable, because it might be difficult to compare characteristics of two points within the same cluster if they have quite dissimilar features.

Algorithm 2 Unified-Weak BinarySearch

Input: Sorted dataset $\hat{\mathcal{S}}_i = \{x_1, \cdots, x_{|\hat{\mathcal{S}}_i|}\}$ in increasing order of $d(x_j, \mu'_p)$, an oracle for weak query Q, target cluster p, set of assignment-known points Z_p, empirical mean μ'_p, and a sampling number $\beta \leq |Z_p|$.

- Search($x_j \in \hat{\mathcal{S}}_i$)**:**
  Select the point x_1 and use it for same-cluster queries
  if $Q(x_1, x_j) = 1$ **then** Set left bound index as $j + 1$
  else if $Q(x_1, x_j) = -1$ **then** Set right bound index as $j - 1$
  else
    Sample $\beta - 1$ points from Z_p. $B \subseteq Z_p$, $|B| = \beta - 1$
    Weak same-cluster query $Q(x_j, y)$, for all $y \in B$
    if x_j is in cluster C_p **then** Set left bound index as $j + 1$
    else Set right bound index as $j - 1$
    end if
  end if
- Stop: Found the smallest index j^* such that x_{j^*} is not in C_p

Output: $r'_i = d(x_{j^*}, \mu'_p)$

Remark 1. Algorithm 2 can also handle oracles with a random behavior. $\beta = 1$ is sufficient for distance-weak oracles.

Local Distance-Weak Oracle We define the first weak-oracle model sensitive to distance, a local distance-weak oracle, in a formal way to include two vague situations described before. These confusing cases for local distance-weak oracle are visually depicted in Figure 1 for better explanation.

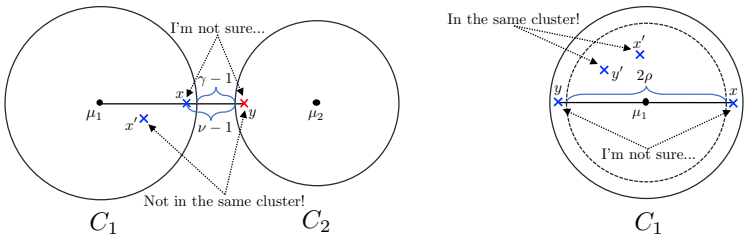


Figure 1: Visual representation of two *not-sure* cases for the local distance-weak oracle. (*Left*) Two points from the different clusters are too close. (*Right*) Two points from the same clusters are too far.

Definition 5 (Local Distance-Weak Oracle)**.** *An oracle having a clustering $\mathcal{C} = \{C_1, \cdots, C_k\}$ for data $\mathcal{X}$ is said to be (ν, ρ) local distance-weak with parameters $\nu \geq 1$ and $\rho \in (0, 1]$, if $Q(x, y) = 0$ for any given two points $x, y \in \mathcal{X}$ satisfying one of the following conditions:*

(a) $d(x, y) < (\nu - 1) \min\{d(x, \mu_i), d(y, \mu_j)\}$, *where* $x \in C_i, y \in C_j, i \neq j$

(b) $d(x, y) > 2\rho r(C_i)$, *where* $x, y \in C_i$

One way to overcome the uncertainty is to provide at least one good point in a query, i.e. *better pairs*. If one of the points x and y for the query $Q(x, y)$ is close enough to the center of a cluster, a local distance-weak oracle does not get confused in answering. This situation is realistic because one

representative data sample of a cluster might be a good baseline when comparing to other elements. Theorem 1 is founded on this intuition, and we show that our modified version of SSAC will succeed if at least one representative sample per cluster is suitable for the weak oracle.

Theorem 1. *If a cluster C_i contains at least one point $x^* \in C_i$ satisfying $d(x^*, \mu_i) < c_{local} \cdot r(C_i)$ for all $i \in [k]$, then combination of Algorithm 1 and 2 outputs the oracle's clustering $\mathcal{C}$ with probability at least $1 - \delta$ by asking weak same-cluster queries to a (ν, ρ) local distance-weak oracle.* $(c_{local} = \min\{2\rho - 1, \gamma - \nu + 1\} - 2\epsilon$, *where* $\epsilon \leq \frac{\gamma - 1}{2})$

Sketch of Proof. We first show the effect of a point close to the center on weak queries. Then the possibility of having a close empirical mean is provided by defining *good sets* and calculating data-driven probability of failure from it. Last, an assignment-known point is identified to remove the uncertainty of same-cluster queries used in the binary search step.

Global Distance-Weak Oracle A global distance-weak oracle fails to answer depending on the distance of each point to its respective cluster center. In this case, both elements x and y should be in the covered range of an oracle if they don't belong to the same cluster.

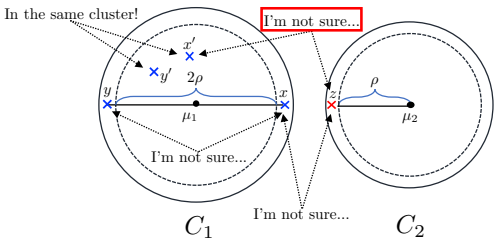


Figure 2: Visual representation of two *not-sure* cases for the global distance-weak oracle. The red box indicates the difference with the local distance-weak oracle.

Definition 6 (Global Distance-Weak Oracle)**.** *An oracle having a clustering $\mathcal{C} = \{C_1, \cdots, C_k\}$ for data $\mathcal{X}$ is said to be ρ global distance-weak with parameter $\rho \in (0, 1]$, if $Q(x, y) = 0$ for any given two points $x, y \in \mathcal{X}$ satisfying one of the following conditions:*

(a) $d(x, \mu_i) > \rho r(C_i)$ *or* $d(y, \mu_j) > \rho r(C_j)$, *where* $x \in C_i, y \in C_j, i \neq j$

(b) $d(x, y) > 2\rho r(C_i)$, *where* $x, y \in C_i$

The problem of a global distance-weak oracle compared to the local distance-weak model is the increased ambiguity in distinguishing elements from different clusters. Nevertheless, once we get a good estimate of the center, better pairs with one good point can be still found to support the oracle in answering same-cluster queries.

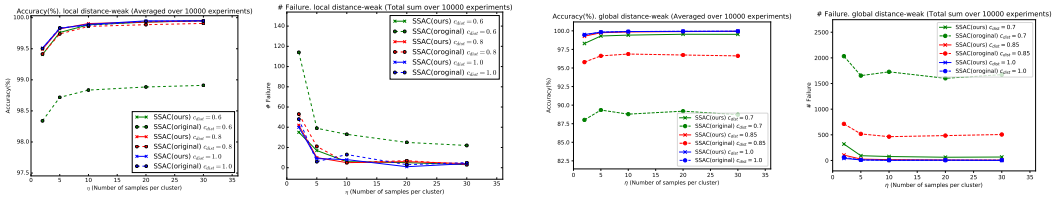
Theorem 2. *If a cluster C_i contains at least one point $x^* \in C_i$ satisfying $d(x^*, \mu_i) < c_{global} \cdot r(C_i)$ for all $i \in [k]$, then combination of Algorithm 1 and 2 outputs the oracle's clustering $\mathcal{C}$ with probability at least $1 - \delta$, by asking weak same-cluster queries to a ρ global distance-weak oracle.* $(c_{global} = 2\rho - 1 - 2\epsilon$, *where* $\epsilon \leq \frac{\gamma - 1}{2})$

4 Experimental Results

Synthetic Data Points of each cluster are generated from isotropic Gaussian distribution. We assume that there exists a ground truth oracle's clustering, and the goal is to recover it where labels are partially provided via weak same-cluster queries. For visual representation, 2-dimensional data points are considered, and other parameters are set to $n = 600$ (number of points), $k = 3$ (number of clusters), and $\sigma_{std} = 2.0$. Data points satisfy γ-margin property with condition $\gamma_{\min} \leq \gamma \leq \gamma_{\max}$. To focus on scenarios with narrow margins, $\gamma_{\min} = 1.0$ and $\gamma_{\max} = 1.1$ are chosen.

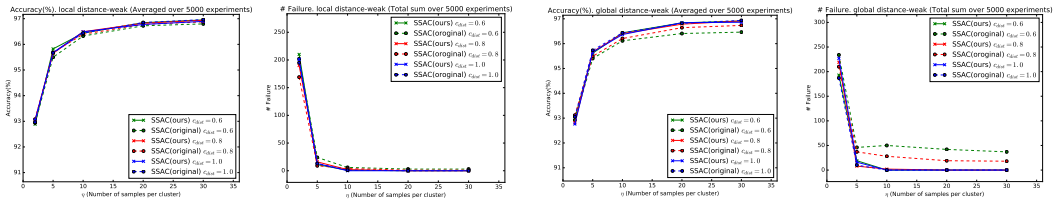
MNIST γ-margin property is difficult to evaluate and satisfy in real world data as a *good* representation or an embedding space is not given. Therefore, we assumed that the oracle has a 2-dimensional embedding space equivalent to the one generated by t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [22]. We used digits 0, 6, and 8 in the subset of MNIST dataset for similarity.[3]

[3]Sample MNIST (2500 points) is from the t-SNE code. `https://lvdmaaten.github.io/tsne/`



(a) $Accuracy$ (%) (b) # $Failure$ (c) $Accuracy$ (%) (d) # $Failure$

Figure 3: Synthetic data. (a),(b): Local distance-weak oracle, $c_{dist} \in \{0.6, 0.8, 1.0\}$. (c),(d): Global distance-weak oracle, $c_{dist} \in \{0.7, 0.85, 1.0\}$. *x-axis*: $\eta \in \{2, 5, 10, 20, 30\}$ (Number of samples)



(a) $Accuracy$ (%) (b) # $Failure$ (c) $Accuracy$ (%) (d) # $Failure$

Figure 4: MNIST. (a),(b): Local distance-weak oracle, $c_{dist} \in \{0.6, 0.8, 1.0\}$. (c),(d): Global distance-weak oracle, $c_{dist} \in \{0.6, 0.8, 1.0\}$. *x-axis*: $\eta \in \{2, 5, 10, 20, 30\}$ (Number of samples)

Evaluation Each round of the evaluation is composed of experiments with different parameter settings on (η, c_{dist}). Parameters for the distance-weak oracles, ρ and ν, are controlled by c_{dist} in the experiments: $\rho = c_{dist}$ and $\nu = \max(1, \gamma) + 2 \cdot (1 - c_{dist})$. β is fixed as 1 since we are only considering distance-weak oracles. η and c_{dist} are varied in each round, and the task is repeated 5000 (MNIST) and 10000 (Synthetic) times. Two evaluation metrics are considered: $Accuracy$ is the ratio of correctly recovered data points averaged over n points, and $\#Failure$ is the total number of failures occurred at cluster-assignments. The best permutation for the cluster labels is investigated based on the distances between estimated centers and true centers for the evaluation. To compare the performance of our improved SSAC, the original one [4] receives random answers, $Q(x, y) = \pm 1$ with probability 0.5, whenever an oracle encounters the case of not-sure. Also, pairs used in the binary search steps are randomly selected from the cluster-known points.

Results An accuracy improves as η increases, and this shows the importance of enough number of samples to succeed in clustering with weak oracles. In fact, even small number of samples are sufficient in practice. Failures of the SSAC algorithm can happen as it is a probabilistic algorithm. When η is really small, the possibility of failure increases as we have only few chances to ask cluster-assignment queries. For example, if $\eta = 2$, only $r = \lceil k\eta \rceil = 6$ points are sampled. Then, if all 6 cluster-assignment queries fail, Phase 1 fails which leads to the recovery of less than k clusters. However, such situations rarely occur if η is large enough.

Results in Figure 3 and 4 show that our improved algorithm (solid lines) outperforms the vanilla SSAC (dashed lines) by allowing not-sure query responses to relax oracles. Especially, results on synthetic data clearly prove the effectiveness of providing better pairs to weak oracles in binary search steps. Our algorithm is robust against the different level of distance weakness. Also, empirical results on MNIST further supports the practicality of our algorithm and weak models.[4]

5 Conclusion and Future Work

This paper presents an approach for utilizing weak oracles in clustering. Specifically, we suggest two realistic types of domain experts who can provide an answer "not-sure" for the same-cluster query. For each model, probabilistic guarantee on discovering the oracle's clustering is provided based on our improved algorithm. In particular, a single element close enough to the cluster center mitigates ambiguous supervision by providing better pairs to an oracle. One interesting future direction is to accommodate embedding learning methods for the real-world clustering tasks.

[4]The source code is available online. `https://github.com/twankim/weaksemi`

References

[1] Nir Ailon, Anup Bhattacharya, Ragesh Jaiswal, and Amit Kumar. Approximate clustering with same-cluster queries. *arXiv preprint arXiv:1704.01862*, 2017.

[2] Hassan Ashtiani and Shai Ben-David. Representation learning for clustering: a statistical framework. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 82–91. AUAI Press, 2015.

[3] Hassan Ashtiani and Ali Ghodsi. A dimension-independent generalization bound for kernel supervised principal component analysis. In *Proceedings of The 1st International Workshop on "Feature Extraction: Modern Questions and Challenges", NIPS*, pages 19–29, 2015.

[4] Hassan Ashtiani, Shrinu Kushagra, and Shai Ben-David. Clustering with same-cluster queries. In *Advances In Neural Information Processing Systems*, pages 3216–3224, 2016.

[5] Pranjal Awasthi, Avrim Blum, and Or Sheffet. Center-based clustering under perturbation stability. *Information Processing Letters*, 112(1):49–54, 2012.

[6] Maria-Florina Balcan and Avrim Blum. Clustering with interactive feedback. In *International Conference on Algorithmic Learning Theory*, pages 316–328. Springer, 2008.

[7] Maria Florina Balcan and Yingyu Liang. Clustering under perturbation resilience. *SIAM Journal on Computing*, 45(1):102–155, 2016.

[8] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

[9] Sugato Basu, Arindam Banerjee, and Raymond Mooney. Semi-supervised clustering by seeding. In *Proceedings of 19th International Conference on Machine Learning*. Citeseer, 2002.

[10] Sugato Basu, Arindam Banerjee, and Raymond J Mooney. Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 333–344. SIAM, 2004.

[11] Sugato Basu, Mikhail Bilenko, and Raymond J Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004.

[12] David Cohn, Rich Caruana, and Andrew McCallum. Semi-supervised clustering with user feedback. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 4(1):17–32, 2003.

[13] Ian Davidson and SS Ravi. Clustering with constraints: Feasibility issues and the k-means algorithm. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 138–149. SIAM, 2005.

[14] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[15] Taewan Kim and Joydeep Ghosh. Semi-supervised active clustering with weak oracles. *arXiv preprint arXiv:1709.03202*, 2017.

[16] Brian Kulis, Sugato Basu, Inderjit Dhillon, and Raymond Mooney. Semi-supervised graph clustering: a kernel approach. *Machine learning*, 74(1):1–22, 2009.

[17] Meena Mahajan, Prajakta Nimbhorkar, and Kasturi Varadarajan. The planar k-means problem is np-hard. In *International Workshop on Algorithms and Computation*, pages 274–285. Springer, 2009.

[18] Arya Mazumdar and Barna Saha. Clustering via crowdsourcing. *arXiv preprint arXiv:1604.01839*, 2016.

[19] Arya Mazumdar and Barna Saha. Query complexity of clustering with side information. In *Advances In Neural Information Processing Systems*, 2017.

[20] Lev Reyzin. Data stability in clustering: A closer look. In *International Conference on Algorithmic Learning Theory*, pages 184–198. Springer, 2012.

[21] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.

[22] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of machine learning research*, 15(1):3221–3245, 2014.

[23] Andrea Vattani. The hardness of k-means clustering in the plane. *Manuscript, accessible at http://cseweb. ucsd. edu/avattani/papers/kmeans_hardness. pdf*, 617, 2009.