Diffusion-based Clustering Analysis of Coherent X-ray Scattering Patterns of Self-assembled Nanoparticles

Hao Huang
Computer Science Dept.
Stony Brook University
haohuang@cs.stonybrook.edu

Shinjae Yoo
Computational Science Center
Brookhaven National Lab.
sjyoo@bnl.gov

Konstantine Kaznatcheev
Photon Sciences
Brookhaven National Lab.
kaznatch@bnl.gov

Kevin G. Yager, Fang Lu
Center for Functional Nanomaterials,
Brookhaven National Lab.
{kyager, flu}@bnl.gov

Dantong Yu, Oleg Gang, Andrei Fluerasu
Brookhaven National Lab.
{dtyu, ogang, fluerasu}@bnl.gov

Hong Qin
Computer Science Dept.
Stony Brook University
qin@cs.stonybrook.edu

ABSTRACT

Coherent X-ray scattering is an emerging technique for measuring structure at the nanoscale. Data management and analysis is becoming a bottleneck in this technique. We present an unsupervised method which can sort and cluster the scattering snapshots, uncovering patterns inherent in the data. Our algorithm operates without resorting to templates, specific noise models, or user-directed learning. We test our methods using scattering images of two-dimensional nanoparticle assemblies. The experimental results show the effectiveness of our algorithm on real world scientific data.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; I.4.9 [**Image Processing and Computer Vision**]: Applications; I.5.4 [**Pattern Recognition**]: Computer Vision

General Terms

Algorithms and Experimentation

Keywords

Nanoparticle Assemblies, AHK, EMD

1. INTRODUCTION

X-ray scattering is a collection of experimental techniques that can quantify structural order at the atomic, molecular, and nano-scale. These techniques consist of directing a high-intensity, collimated beam of X-rays at a sample of interest, and measuring the intensity of scattered X-rays as a function of angle; typically through the use of two-dimensional detectors. These scattering "images" non-trivially encode the details of structural order. Coherent X-ray scattering is a variant wherein a small coherence beam is scanned over a sample surface; the coherent scattering image encodes the local ordering illuminated by the beam. As X-ray technology improves, modern instruments are able to generate data at an ever-increasing pace.

Manually clustering/discovering new patterns through the large sets of snapshots is extremely time-consuming; in many cases the wealth and diversity in datasets makes it impossible for a human experimenter to rigorously analyze. Therefore it is necessary to have an unsupervised and robust clustering algorithm to automatically solve the pattern discovery problem across these datasets. The clustering algorithm needs to solve the following challenges when applying usual statistical techniques to X-ray scattering snapshots: 1) it remains unknown what kind of particular features are informative for the clustering task; 2) under X-ray illumination, assembly projection may rotate, therefore the snapshots may provide different slices through three-dimensional spaces; and 3) the snapshots may be corrupted by detectors artifacts, such as image saturation and background burst.

In this paper, we propose an unsupervised, effective and stable algorithm for clustering the scattering snapshots, with the following contributions:

1. Our approach is based on spectral clustering (Section 2) [10] [9], which captures nonlinear correlations within a dataset to cluster the scattering snapshots.

2. We explore the Earth Mover's Distance (EMD) [13] which has desired properties in revealing snapshot distance, and integrate it into our clustering algorithm. We combine EMD with Gaussian kernel, called Earth Mover's Similarity (EMS, Section 3), which allows for partial matching that is highly effective to deal with 2D snapshot occlusions and clutters.

3. Particularly, we apply an advanced spectral clustering, called Aggregated Heat Kernel (AHK, Section 4) [6], which can maximize the intra-cluster similarity while avoiding the influence from noise and artifacts.

4. Our novel algorithms for 2D scattering pattern anal-

SAC'14 March 24-28, 2014, Gyeongju, Korea.
Copyright 2014 ACM 978-1-4503-2469-4/14/03 ...$15.00.
http://dx.doi.org/10.1145/2554850.2554914

ysis combines AHK and EMD in a systematic framework (Section 5).

5. The proposed framework does not rely on any template, specific noise model, or user-directed learning.

6. Experimental results (Section 6) show that our algorithm can effectively sort and cluster the scattering snapshots from nanoparticle assembly.

2. SPECTRAL CLUSTERING

```
Algorithm 1: SpectralClustering($X$, $c$)
  Input: $X \in R^{n \times m}$ where $n$ is the
         #instances(snapshots), $m$ is #attributes, and $c$
         is #clusters.
  Output: Cluster assignments of $n$ instances.
1 Construct the affinity matrix $W \in R^{n \times n}$ using
  $W_{(GAU)}(i,j) = exp(-||x(i) - x(j)||/2\sigma^2)$ ;
2 Compute the diagonal matrix $D \in R^{n \times n}$ where
  $D(i,i) = \sum_{j=1}^{n} W(i,j)$ and $D(i,j) = 0$ if $i \neq j$ ;
3 Compute the graph Laplacian $L$ with $L_{nn} = D - W$,
  $L_{rw} = I - D^{-1}W$ or $L_{sym} = I - D^{-1/2}WD^{-1/2}$ ;
4 Compute the first $c$ nontrivial eigenvectors $\psi$ of $L$,
  $\psi = \{\psi_1, \psi_2, \dots, \psi_c\}$ ;
5 Re-normalize the rows of $\psi \in R^{n \times c}$ into
  $Y_i(j) = \psi_i(j)/(\sum_l \psi_i(l)^2)^{1/2}$ ;
6 Run $k$-means with $c$ and $Y \in R^{n \times c}$.
```

Among a variety of clustering algorithms, we focus on spectral clustering (Algorithm 1), which gained popularity in the last decade in the data mining community because of its ability to discover embedded data structures. Spectral clustering has a strong connection with graph cut, i.e., it uses eigenspace to solve the relaxed forms of the balanced graph partitioning problem [10]. Another advantage is that it can capture the nonlinear manifold structure, which is difficult for many other cluster algorithms, such as k-means [5] and the other linear methods.

In 2011, Yoon et.al applied spectral clustering on single-particle X-ray diffraction snapshots [16]. Each of the snapshots/instances in spectral clustering is represented by a feature vector $x(i) \in R^{1 \times m}$. The final vectors Y represent the global manifold structure of the entire snapshot dataset. Here spectral clustering exploits similarity between snapshots to discover the embedded dimensions. Therefore, the embedded structure captured in Step 5 (normalized eigenvectors) is highly dependent on the affinity matrix constructed in Step 1 and Step 2, and the normalization in Step 3.

However, there are three challenges with similarity measurement on the 2D nanoparticle assembly snapshots:

1. It is not easy to find the informative features to evaluate similarity between two different snapshots;

2. Those snapshots containing noise and artifacts, and the selection of scaling parameter σ in Gaussian kernel (Step 1 in Algorithm 1) could affect clustering results radically because of their influence on the neighborhood information in the affinity matrix;

3. The sampling density of each type of nanoparticle assembly could be different, which could incur negative effects across clusters during clustering.

In our research, we explore the use of two advanced techniques to solve these challenges: Earth Mover's Distance (to solve challenge 1), and Aggregated Heat Kernel (to solve challenge 2 and 3).

3. EARTH MOVER'S SIMILARITY

Earth mover's distance (EMD) [13] was introduced in computer vision and applied in many other fields as a highly adaptable distance measurement that can be tweaked to closely match human perception [1]. It interprets a feature vector as a distribution of earth which needs to be transformed (or moved) to another distribution of earth (i.e. feature vector). The reported distance between two feature vectors is derived from the minimum mapping of one vector to the other, where each mapping is measured in terms of the amount of transported earth multiplied by the unit cost of moving from its source to its destination [1].

In other words, EMD considers not only feature difference in those matching dimensions - as many other distance metrics such as Euclidean distance would do - but also the difference between non-matching dimensions. This gives EMD an advantage as EMD allows for partial matching in a very natural way, which is of importance in dealing with the occlusions and clutter that occur in scattering snapshots. Therefore, by using EMD we can measure variance between the snapshots of nanoparticle assembly, without manually extracting specific features.

Computing EMD can be formalized as a linear programming problem [13]: denote the first snapshot with m rows as $P = (p_1, w_{p_1}), ..., (p_m, w_{p_m})$, where p_i is the row representation and w_{p_i} is the weight of the row; the second snapshot $Q = (q_1, w_{q_1}), ..., (q_n, w_{q_n})$ is denoted in the same way. $\mathbf{D} = [d(i,j)]$ is the ground distance matrix where $d(i,j)$ is the ground distance between rows p_i and q_j. In our implementation we apply 2-norm distance as ground distance. EMD needs to find a flow $\mathbf{F} = [f(i,j)]$, with $f(i,j)$ the flow variable between p_i and q_j that minimizes the overall cost:

$$\text{WORK}(P, Q, \mathbf{F}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f(i,j) d(i,j) \ ,$$

which subjects to the following constraints:

$$\begin{aligned}
f(i,j) &\geq 0, && 1 \leq i \leq m,\ 1 \leq j \leq n \\
\sum_{j=1}^{n} f(i,j) &\leq w_{p_i}, && 1 \leq i \leq m \\
\sum_{i=1}^{m} f(i,j) &\leq w_{q_j}, && 1 \leq j \leq n \\
\sum_{i=1}^{m} \sum_{j=1}^{n} f(i,j) &= \min(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}) \ .
\end{aligned}$$

According to the analysis in [13], the first constraint allows moving "supplies" from P to Q and not vice versa. The next two constraints limit the amount of supplies sent by the rows in P to be less than their weights w_{p_i}, and allow no more supplies received by the rows in Q than their weights w_{q_i}.

The last constraint forces to move the maximum amount of supplies possible. This amount is called "total flow". Once the transportation problem is solved with the solution of the optimal flow $\mathbf{F}$, the EMD can be subsequently defined as the total cost normalized by the total flow:

$$\mathrm{E}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f(i,j)d(i,j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} f(i,j)} \ , \qquad (1)$$

where the normalization factor in the denominator is introduced in order to obtain stable results.

However, we need to use similarity instead of distance measurement in spectral clustering. Therefore we combine EMD into the normal Gaussian kernel, and derive a new kernel, which we call Earth Mover's Similarity (EMS):

$$W_{(EMS)}(P,Q) = exp(\frac{-\mathrm{E}(P,Q)}{2\sigma^2}) \ . \qquad (2)$$

Although EMS matches the human perception (i.e. perceptual similarity) of images better than other similarity measurements, we still need a more "manifold-aware" and "density-aware" approach to assemble the snapshots with similar nanoparticle assembly structures, but with different rotation angles. In the next section, we introduce Aggregated Heat Kernel [6] to achieve these goals.

4. AGGREGATED HEAT KERNEL

In this section, we integrate Green's function and Laplace-Beltrami Normalization into our framework to achieve better manifold-aware and density-aware properties, especially for the snapshots with similar nanoparticle assembly structure but different rotation angles. Specifically we apply Aggregated Heat Kernel (AHK) [6] in the research.

4.1 Green's Function

Green's function[3] [15] is an infinite time scale analysis. It can be derived from heat kernel through applying infinite integral along the entire time dimensions. Huang et al. proposed AHK [6] that is built upon Green's function, and applied it to design a more robust spectral clustering. The kernel function used in [6] is defined as:

$$W_{(AHK)}(i,j) = \sum_{k} \left[\frac{1}{\lambda_k + \gamma} \psi_k(i)\psi_k(j) \right] , \qquad (3)$$

where ψ and λ are the eigenvectors and eigenvalues extracted from the graph Laplacian $L = \psi'\lambda\psi$ which derived from the (normalized) affinity matrix W, and γ is a smoothing factor. The most significant benefit of Equation 3 is that it takes all possible paths with the entire time dimensions into consideration [6]. Specifically it has great potential to connect the snapshots with similar nanoparticle assembly structure together.

4.2 Laplace-Beltrami Normalization

It is important to find the best way of graph Laplacian for AHK. It is shown in [8] that if we assume uniform sampling of data points from a sub-manifold $\mathcal{M}$, the eigenvectors of L_{rw} with $\sigma \to 0$ and $n \to \infty$ tend to approximate the Laplace-Beltrami operator on $\mathcal{M}$ and guarantee to reconstruct the manifold structure. However, in reality, data samples are inclined to be non-uniform and show skewed density distributions, resulting in a poor manifold reconstruction in AHK. To mitigate the distribution sensitivity of random walk (RW) normalization, the following two additional normalizations are considered:

$$W^{(\alpha)} = D^{-\alpha} W D^{-\alpha}, \qquad (4)$$

$$L^{(\alpha)} = I - {D^{(\alpha)}}^{-1} W^{(\alpha)}, \qquad (5)$$

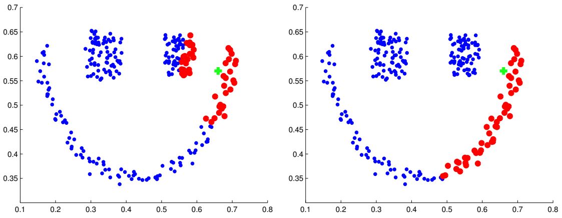
where α is a normalization factor and $D^{(\alpha)}$ is a diagonal matrix with the sum of row weight of $W^{(\alpha)}$.

- If $\alpha = 0$, $L^{(0)} = L_{rw}$ (random walk normalization).
- If $\alpha = 1/2$, then it is *Fokker-Planck* (FP) diffusion.
- If $\alpha = 1$, it is Laplace-Beltrami normalization (LBN).

The relations among these three normalizations are well described in [4]. Depending on α, LBN can also be reduced to RW or FP diffusion. In particular, we focus on LBN because it removes the influence of dataset density and recovers manifold structures on $\mathcal{M}$ with the condition of both $\sigma \to 0$ and $n \to \infty$ [4]. In other words, the additional re-normalization of W enables us to reconstruct manifold structures even under non-uniform density distribution. As a result, our clustering results can also be less sensitive to noise, artifacts, and scaling parameter such as σ.

4.3 Properties of AHK

It is worth to notice that AHK has the following contributions: 1) It is more robust and less sensitive to noise or artifacts than other regular kernels. 2) To mitigate the biased contribution of the denominator from some extremely small eigenvalues λ, AHK introduces a smoothing term γ in Equation 3 to make the computation more stable. 3) To relieve the non-uniform density distribution effect, AHK employs Laplace-Beltrami normalization (LBN) [6] which can remove the influence of dataset density and recover the Riemannian manifold under skewed density distribution. In other words, AHK enables better and more stable manifold reconstruction, especially under noise, parameter disturbance, and non-uniform density distribution. Therefore in theory it guarantees the strong adjacency (similarity) between snapshots with similar nanoparticle assemblies.



(a)

(b)

Figure 1: Comparison of the 50 nearest neighbor retrieval (in red) of the green cross point using Gaussian kernel (1(a)) and AHK (1(b)). Obviously AHK shows better manifold-aware property.

Figure 1 shows a side-by-side comparison of 50 nearest neighbor retrieval of the green cross point with Gaussian kernel (Step 1 in Algorithm 1) and AHK (Equation 3). We can observe that AHK performs better in capturing the unobserved manifold structure of Gaussian kernel. This results come from not only the well reconstructed manifold by Equation 3, but also the relaxation of non-uniform density distribution with the proper LBN.

5. EMS-AHK CLUSTERING ALGORITHM

```
Algorithm 2: EMS-AHK($X$, $c$)
  Input: $X$ is a collection of $n$ snapshots and $c$ is #clusters.
  Output: Cluster assignments of $n$ snapshots.
1 Construct the affinity matrix $W_{(EMS)} \in R^{n \times n}$ using Equation 2 ;
2 Perform Laplace-Beltrami normalization $L_{\mathrm{LBN}}$ on $W_{(EMS)}$;
3 Compute generalized eigenvectors $\psi_i$ and the corresponding eigenvalues $\lambda_i$, $i = 1, 2, ..., n$ ;
4 Construct $W_{(AHK)}$ matrix with $\psi_i$ and $\lambda_i$ using Equation 3 with $\gamma = 0.001$ ;
5 Compute the first nontrivial $c$ eigenvectors $\Psi$ of $W_{(AHK)}$, $\Psi = \{\Psi_1, \Psi_2, \ldots, \Psi_c\}$;
6 Run $k$-means with $c$ and $\Psi \in R^{n \times c}$.
```

We describe our algorithm framework in Algorithm 2. First of all, it computes the similarity between each snapshot pair by using EMS (Step 1). Second, it undergoes a data warping process by using LBN and AHK (Step 2-4). Then we perform eigen-decomposition (Step 5) and construct the embedding projection (Step 6). K-means algorithm is used in the last step (Step 7) for final clustering results.

We select three typical snapshots of nanoparticle assembly, and rotate each of them with 10 different angles. Figure 2(a), 2(b) and 2(c) show some of the three types of snapshots with different angles. Figure 2(d) shows different ways of embedding reconstruction. The projection derived from NJW [10] using Gaussian kernel (Figure 2(d1), Gau-NJW) apparently organizes the snapshots with the same pattern within a circular shape. Comparably, Gaussian kernel with subsequent AHK (Figure 2(d2), Gau-AHK) provides more compact embedding structure w.r.t. Gau-NJW. However, it is hard to separate the red and green clusters of snapshot. EMS is aware of the difference between the circular-shaped red and green cluster. Hence in Figure 2(d3) (EMS-NJW), the three clusters have clear separation between each other. Nonetheless, EMS incorrectly amplifies the differences of snapshots generated from the same pattern but with different rotation angles, and breaks one class into two separated classes. AHK statistically depicts the traces of all random walk, and thereby has an intrinsic potential to assemble all the similar snapshots with angles that continuously change. Therefore it is shown in Figure 2(d4) that, EMS-AHK is the best choice for clustering coherent X-ray scattering snapshots, which maximizes the intra-cluster similarity while minimizing the inter-cluster similarity.

6. EXPERIMENTAL ANALYSIS

6.1 Data Set Generation and Preprocessing

The soft X-ray scattering measurements were performed at the Canadian Light Source scanning transmission X-ray microscope (STXM) [7]. The 700 eV photon energy with monochromaticity of 5,000 resolving power was chosen as the best compromise between flux delivered to sample, high efficiency of focusing optics Fresnel zone plate (FZP) with 45 nm outermost zone width, negligible parasitic scattering through optical sorting aperture and desired sample-to-optics distance. The spot size at the sample position was increased to 200 nm (so that several nanoparticles were illuminated at once). The sample cell consisted of thin $\mathrm{Si_3N_4}$ windows (50 nm thickness, and frame size of 250 μm) coated with dried nanoparticle dispersion. This cell was set perpendicular to the beam at a short distance behind the zone plate focus. An in-vacuum back-illuminated CCD camera (Andor DX) was placed 40 mm behind the sample, covering an angular range up to a reciprocal vector of 1 nm^{-1}. To extend the camera dynamic range exposures with short (10 ms) and long (200 ms) dwell times were stitched to provide a single scattering image with appreciable scattering intensity extending to a reciprocal vector of 0.4 nm^{-1}. The sample was raster-scanned through the X-ray spot with 100 nm steps so that illuminated areas of adjacent acquisition positions are overlapped. A typical scan covers a 2×2 μm portion of the samples. Several representative regions were measured.

The dataset analyzed consisted of 3,778 scattering snapshots. Our aim is to cluster these snapshots and discover useful patterns of nanoparticle assembly. Prior to analysis, we have three steps of preprocessing: 1) normalize snapshots by dividing each pixel with the maximum intensity of the snapshot; 2) enhance snapshot contrast by transforming the values of intensity, so that the histogram of the output snapshot approximately matches a specified (default) histogram; 3) downsample snapshots to 200×200 pixels (still sufficient so that each individual speckle in the image covers several pixels) in order to keep computational time reasonable while computing EMD.

6.2 Experiment Details

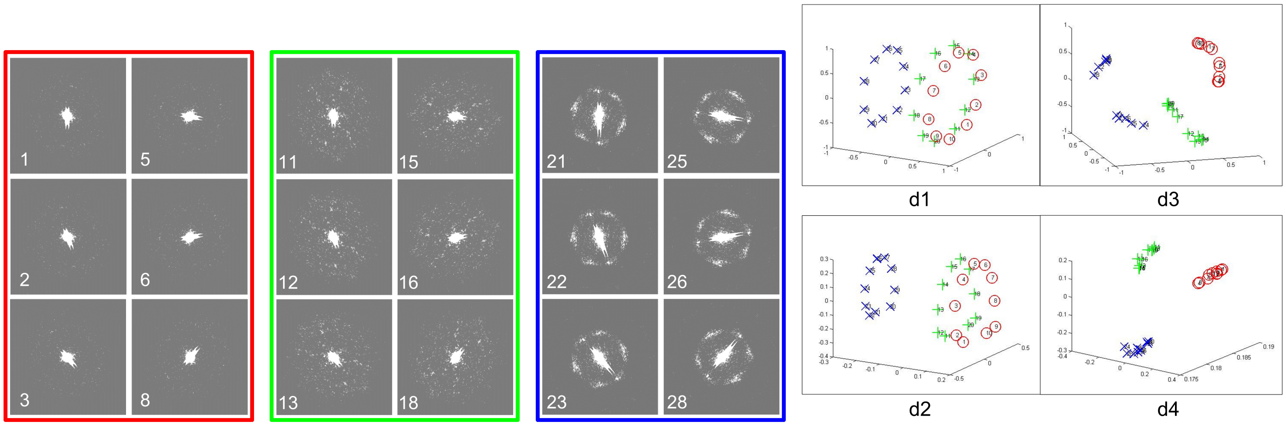
Even though in our implementation we adopted a fast version of EMD [12], the computation is still time-consuming. To make the algorithm more efficient, we only compute EMD within k nearest neighbor (k-nn) with $k = 20$ in a Euclidean space. In other word, for each snapshot we only compute the EMS to the 20 nearest snapshots in Euclidean space. The k-nn affinity construction then creates 36,663 edges.

EMS (Equation 2) was used to construct the affinity matrix. To control the scaling parameter adaptively and at the same time preserve local density information, we compute the average EMD between each snapshot to its q-nearest neighbors, and use this average EMD (noted as σ_q) to set the scaling parameter σ in Equation 2. In our experiments we set $q = 10$.

After LBN and AHK construction, we plot the first three nontrivial eigenvectors as shown in Figure 3. According to the projected space distribution, we set the number of clusters/patterns to be 8. To partition data into discrete clusters, k-means clustering was applied and each cluster was labeled with different color. Of the total 3,778 snapshots, black cluster has 1109 snapshots, yellow cluster has 679 snapshots, purple cluster has 493 snapshots, brown cluster has 458 snapshots, light blue has 356 snapshots, dark blue has 257 snapshots, red cluster has 218 snapshots, while green cluster has 208 snapshots.

6.3 Evaluation Metric

In order to evaluate the reliability of the clustering results, we manually labeled 200 randomly selected snapshots. Each snapshot has only one label according to their visual similarity to the 8 patterns we found in Figure 3. We use these labels as ground truth for our result evaluation.



(a) Type 1 (b) Type 2 (c) Type 3 (d) Four different Embedding Projections

Figure 2: Three typical snapshots of nanoparticle assembly, marked as red (2(a)), green (2(b)) and blue (2(c)) respectively. Four different ways of embedding projection are shown in 2($d1$) (by Gau-NJW), 2($d2$) (by Gau-AHK), 2($d3$) (by EMS-NJW), and 2($d4$) (by EMS-AHK). Obviously EMS-AHK provides more cluster-aware embedding structure.

Since now we have the ground truth labels for the 200 snapshot, we compare our clustered results with these labels. We use several widely used evaluation metrics in our experiment (e.g., purity, normalized mutual information (NMI)). Due to space limitation, NMI is used as our only evaluation metric listed in this paper because most of the clustering algorithm papers make use of NMI as their primary evaluation metric. The definition of NMI can be referred to [14].

6.4 Results and Analysis

We compare our algorithm (EMS-AHK) to three competitive algorithms: 1) NJW [10] with Gaussian kernel using Euclidean distance (Gau-NJW [16]); 2) AHK [6] with Gaussian kernel using Euclidean distance (Gau-AHK); 3) NJW with EMS kernel (EMS-NJW). Table 1 summarizes the NMI score comparisons, which once again, confirms our observation in the previous experiments in Figure 2:

1. The difference between Gau-NJW and Gau-AHK shows that AHK raises the clustering performance 17%. It substantiates that in this application, AHK maximizes the intra-cluster similarity while minimizing the negative influence from noise and artifacts of snapshots.

2. The comparison between Gau-NJW and EMS-NJW demonstrates that EMS improves the clustering performance 122%, which means that EMS has desired properties in revealing the intrinsic similarity of coherent X-ray scattering snapshots, without manually selecting informative features.

3. The combination of AHK and EMS, boosts the clustering performance 150%. Thereby our novel algorithm can effectively sort and cluster scattering snapshots, uncovering patterns inherent in the data.

The computational cost of the algorithm is determined primarily by EMD and eigen-decomposition. The fast implementation of EMD has $O(m^2 \log m)$ time complexity [12], where m is the number of bins in a single snapshot. In our implementation, the k-nn approach produces sparse matrices, and the subsequent steps requires less CPU time for computing EMD and performing eigen-decomposition. There are many iterative methods to conduct eigenvalue decomposition (e.g., power iteration [2]), but in general solving the eigen-decomposition can be reduced to matrix multiplication by computing a symbolic determinant, which has a running time of $O(n^3+n^2\log^2 n)$ [11], where n is the number of snapshots.

7. CONCLUSION

We have presented an unbiased, stable, practical and effective algorithm for clustering/discovering new patterns through the experimental 2D snapshots, which are produced by a focused X-ray beam to probe the local coordination of particles within the nanoparticle assembly. Our proposed algorithm is unsupervised without resorting to user-directed learning, specific noise models, or templates. Therefore it is an essential step in realizing the full potential of the possibility demonstrated by recent results on nanoparticles. Immediate future work will be concentrated on improving the efficiency and accuracy, and constructing local and global coordinates with the goal of learning the embedding structure of the snapshot dataset.

8. ACKNOWLEDGEMENT

Research carried at Center for Functional Nanomaterials, Brookhaven National Laboratory, supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886. This research is also a part of “A NSLS-II Workflow Prototype System For Supporting Data Intensive Beamline Experiments”, and supported in part by LDRD No. 13-017.

9. REFERENCES

[1] I. Assent, M. Wichterich, T. Meisen, and T. Seidl. Efficient similarity search using the earth mover’s distance for large multimedia databases. *ICDE*, 2008.

[2] R. Badeau, B. David, and G. Richard. Fast approximated power iteration subspace tracking. *IEEE Signal Processing*, pages 2931–2941, 2005.

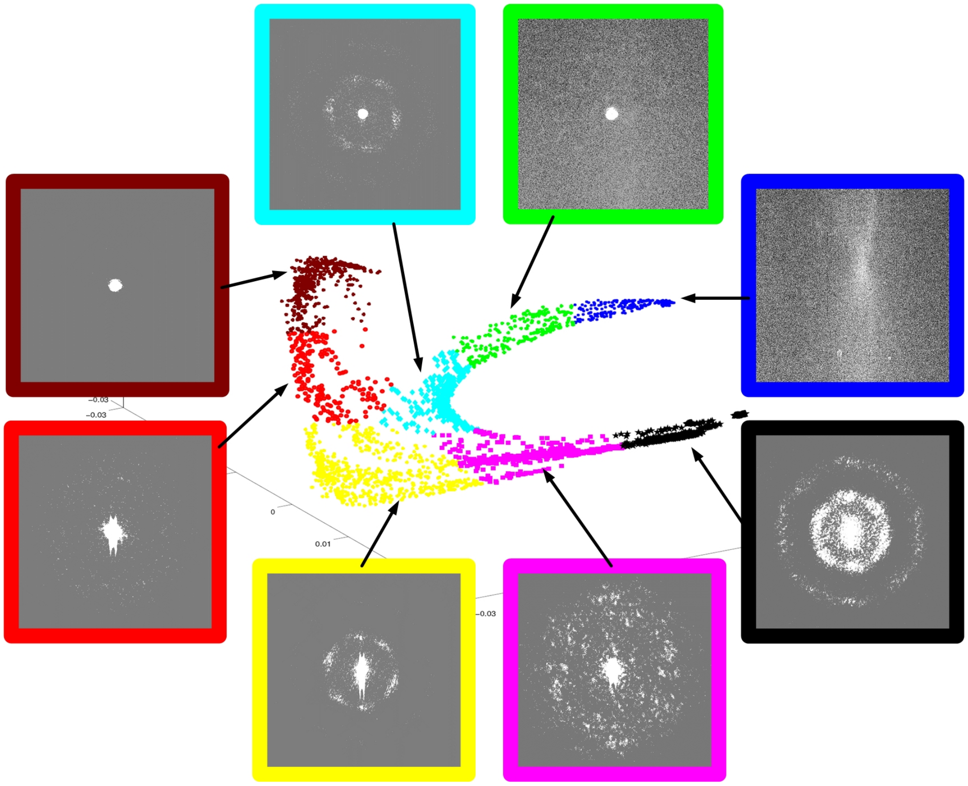


Figure 3: The first three nontrivial eigenvectors of EMS-AHK projection.

Table 1: Statistics of Experiment Comparisons

Algorithms	Gau-NJW	Gau-AHK	EMS-NJW	EMS-AHK
NMI	0.3543	0.4138	0.7869	0.8872

[3] S. Y. Cheng, P. Li, and S. T. Yau. Heat equations on minimal submanifolds and their applications. *American Journal of Mathematics*, (5):1033–1065, 1984.

[4] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[5] J. A. Hartigan and M. A. Wong. Algorithm as 136: a k-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1978.

[6] H. Huang, S. Yoo, H. Qin, and D. Yu. A robust clustering algorithm based on aggregated heat kernel mapping. *IEEE ICDM*, pages 270–279, 2011.

[7] K. V. Kaznatcheev, C. Karunakaran, U. D. Lanke, S. G. U. amd M. Obst, and A. P. Hitchcock. Soft x-ray spectromicroscopy beamline at the cls: Commissioning results. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 582(1):96–99, 2007.

[8] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multi-cue data matching by diffusion maps. *IEEE TPAMI*, 28(11):1784–1797, 2006.

[9] U. V. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[10] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14:846–856, 2002.

[11] V. Y. Pan and Z. Q. Chen. The complexity of the matrix eigenproblem. *ACM STOC*, pages 507–516, 1999.

[12] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV*, 2009.

[13] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[14] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, pages 583–617, 2003.

[15] S. T. Yau. Open problems in geometry. *Proc. Symp. Pure Math*, 1993.

[16] C. H. Yoon, P. Schwander, C. Abergel, and et.al. Unsupervised classification of single-particle x-ray diffraction snapshots by spectral clustering. *Optics Express*, 19(17):16542–16549, 2011.