The Structure of Halos: FOF vs. DBSCAN

- The dark matter halo mass function is a key repository of cosmological information over a wide range of mass scales, from individual galaxies to galaxy clusters. N-body simulation shows that Friend-of-Friend (FOF) mass function has a universal form to a surprising level of accuracy. However, observed group and cluster masses are usually stated in terms of a spherical over-density (SO) mass, which does not map simply to the FOF mass. Results from Monte Carlo realizations of ideal Navarro-Frenk-White (NFW) halos and N-body simulation show that FOF and SO map 80-85% halos only if concentrations are known.
- Challenge: Bridged halos complicates the mapping between FOF halo and SO halo.
- Solution: Investigating the mapping between DBSCAN halo and SO halo: Contrast the properties of DBSCAN with FOF; Investigating relation of DBSCAN to percolation theory similar to FOF; and Investigating whether over-linking problem of FOF can be mitigated by DBSCAN.
- Experiment: DBSCAN with 100,000 Monte-Carlo samples, 1,000 particles per sample, c = 5.
- Results: NFW and DBSCAN mass ratio is close to 1, amplitude is high, and deviation range is smaller.

Multiple Instance Learning (MIL)

- Problem being solved
  - Land-use/land-cover classification using very high-resolution (VHR) remote sensing imagery
- State of the art
  - Single instance learning algorithms (statistical, decision trees, neural networks, …) are not efficient for recognizing complex patterns in VHR images
  - MIL approaches like Citation-KNN are computationally expensive

Bag of Gaussian MIL (BoG MIL)

- BoG MIL is a novel computationally efficient algorithm
- Models all instances in a segment as a Gaussian distribution
- Each land use/land cover is modeled as bag of Gaussian (as opposed to single Gaussian per class)
- Prediction is based on statistical matching and ranking

Accuracy:

<table>
<thead>
<tr>
<th>City</th>
<th>Citation-KNN</th>
<th>Regression</th>
<th>RF</th>
<th>MLP</th>
<th>NB</th>
<th>BoG MIL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acrra</td>
<td>76.25</td>
<td>71.25</td>
<td>72.08</td>
<td>69.58</td>
<td>75.66</td>
<td>95.66</td>
</tr>
<tr>
<td>Caracas</td>
<td>82.96</td>
<td>78.15</td>
<td>81.85</td>
<td>81.81</td>
<td>74.07</td>
<td>85</td>
</tr>
<tr>
<td>La Paz</td>
<td>80.97</td>
<td>77.17</td>
<td>78.26</td>
<td>80.23</td>
<td>76.08</td>
<td>83.25</td>
</tr>
<tr>
<td>Kandahar</td>
<td>79.78</td>
<td>64.89</td>
<td>69.14</td>
<td>73.93</td>
<td>60.1</td>
<td>81.2</td>
</tr>
</tbody>
</table>

Distance Field Based Analysis & Visualization

- Computing distance fields is a fundamental requirement for many algorithms of data visualization and analysis.
- We have designed and implemented a new spatial data structure, named parallel distance tree, to enable highly scalable parallel distance field computing.
- The method is general to support various data types (including, but not limited to, polygonal objects, point/ particle data, and volumetric data) and handle different distance metrics (including, but not limited to, Euclidean distance, City block distance, and Chessboard distance).
- We have integrated our method with real-world large scientific simulations to support in-situ processing and data reduction.
- The design does not depend on any particular architectures, and the scalability has been demonstrated on state-of-the-art supercomputers.
- The resulting technology will benefit many application areas from fusion, combustion, to climate, and astrophysics simulations.

STPMiner

- STPMiner is an high-performance spatiotemporal pattern mining toolbox for analyzing big spatiotemporal datasets.
- Solution: It offers computationally efficient data mining primitives tailored for heterogeneous architectures.
- Applications: Spatial classification (Land use/land cover mapping), clustering (earth science), change detection (biomass monitoring), and co-location pattern detection (climate change impacts).