Scientific Data Mining

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Dept. of Computer Science, Dept. of Physics & Astronomy,
Dept. of Chemical Engineering & Materials Science,
Department of Biological Sciences
University of Southern California

Email: anakano@usc.edu
Scientific Data Mining

- **Scientific data mining**: Automated detection of knowledge hidden in large & often noisy scientific (experimental, simulation, etc.) datasets
- **Knowledge**: Simplest (*i.e.*, minimal description length) explanation to replace exhaustive enumeration of the original data

\[ m \frac{d^2}{dt^2} \ddot{r}(t) = \bar{F} \]
Parallel computing on globally distributed supercomputers & visualization platforms will revolutionize & democratize science & engineering (*e.g.*, Google astronomy in the flat world)

**Google Science in the Flat World**

Gliese 570D (plus companion)  
*Artist’s Rendition*

SDSS image of brown dwarf,  
2MASSI J0104075-005328  
http://www.us-vo.org
Atomistic Data as a Graph

• Molecular dynamics data
  — Atomic data: species, positions, velocities, stresses,…
    \[ \{ \lambda_i, \vec{r}_i, \vec{v}_i, \sigma_i, \ldots \mid i = 1, \ldots, N \} \]
  — Atomic-pair data: bond order, pair distance,…
    \[ \{ B_{ij}, \vec{r}_{ij}, \ldots \mid i, j = 1, \ldots, N; i \neq j \} \]

• Chemical bond network
  \[ G = (V, E) \]
  \[ V: \text{Set of atoms} \]
  \[ E: \text{Set of bonds} \]

  — Node degrees
  — Paths
  — Rings
  — Frequently occurring subgraphs
Hypervelocity Impact on Ceramics

- 209M-atom MD of AlN
- 300M-atom MD of SiC
- 540M-atom MD of Al₂O₃

↑ [0001]

- Al₂O₃ plate
- 18 km/s impact

0.2 µm
Shock-Induced Structural Phase Transformation in AlN

- Wurtzite (4-coordinated) to rocksalt (6-coordinated) phase transformation at 20 GPa
Stress Domains in $\text{Si}_3\text{N}_4$/Si Nanopixels

Stress well in Si with a crystalline $\text{Si}_3\text{N}_4$ film due to lattice mismatch

Stress domains in Si due to an amorphous $\text{Si}_3\text{N}_4$ film
Si(111)/Si$_3$N$_4$(0001) Interface
Stress Domains in Si/Si$_3$N$_4$ Nanopixel

Stress [GPa]

-1 0 1

Si(111)/Si$_3$N$_4$(0001) interfaces

I

Si$_3$N$_4$: N  Si

Si:

II

Cross-Cut Degree!

Lattice mismatch (1%) induced interfacial domain array

N-Si coordination across interface

1 2 3

50nm

Belk et al. ('97)

Misfit dislocation network in InAs/GaAs(111)
High-Pressure Structural Transformation

- Wurzite (node degree 4) to rocksalt (node degree 6) structural transformation of a GaAs nanoparticle under high pressure

- Existence of multiple domains?

2.5 GPa
30 Å
123 Å

~20 GPa
node degree 4
node degree 6
Graph-Transition Tracking

- Finite set of graph transitions as a classifier

\[ G = (V, E) \]
\[ G' = (V, E') \]
\[ E \subseteq E' \]

Graph Transition!
Chemical Reaction Network


Zhang et al., Science 318, 1121 ('07)

Yin et al., Nature 451, 318 ('08)

Arnold group, Nature Rev. MCB 10, 867('09); COCB 13, 3 ('09)

Chen et al., Nature Nanotechnol. 8, 755 ('13)

Reaction graph = language for self-assembly & catalytic cycle design

Directed & accelerated evolution
- Oxide thickness saturates at 40 Å after 0.5 ns, in agreement with experiments
- Oxide region/metal core is under negative/positive pressure
- Attractive Al-O Coulomb forces contribute large negative pressure in the oxide
Oxidative Percolation

Clusters of OAl₄ coalesce to form a neutral, percolating tetrahedral network that impedes further growth of the oxide.

_Percorative Connected Components_!
Fractal Nanocarbon Product

- Percolation transition causes carbon clusters to exhibit power-law distribution of sizes: $C(i) \sim i^{-\tau}$

- Fractal nanocarbon product with large surface areas may find supercapacitor, battery-electrode & mechanical metamaterial applications: $d_f = d/(\tau - 1) \sim 1.85$

K. Nomura et al., Sci. Rep. 6, 24109 ('16)

J. Insley et al., IEEE/ACM SC16
Shortest-Path Rings

- **K-ring:** Given a vertex $x$ & two of its neighbors $w$ & $y$, a K-ring generated by the triplet $w-x-y$ is any ring containing the edges $[w-x], [x-y]$ and a shortest path $w-y$ path in $G-x$
Ring-based Data Mining

Shortest-path ring analysis of intermediate-range order (IRO) in disordered materials

Correlation between IRO in neutron scattering & ring distribution
Fast Ring Analysis: Dual-Tree Expansion

Original Algorithm

Proposed Algorithm

C. Zhang et al., Computer Physics Communications 175, 339 ('06)
DTE Algorithm

Algorithm dual_tree_expansion()

Input:
- \( V \) = Set of all vertices (i.e., atoms)
- \( R_c \) = Ring cutoff range (Euclidean)
- \( R_b \) = Bond cutoff distance (Euclidean)
- \( L_{\text{max}} \) = Maximum length of ring (integer)
- \( P \) = Number of compute nodes

Output:
The K-ring statistics for all vertices in the network
List of atoms with abnormal ring profile

Variables:
- \( \text{Neighbors}(V) \) = Set of vertices that share an edge with vertex \( V \)
- \( K_p(V) \) = Number of \( p \)-member rings that go through vertex \( V \)
- \( L_v \) = Length of the ring formed with path \( (V_0, V, V) \)

Steps:
0 coarse grained spatial decomposition of atoms on \( P \) compute nodes with a thin boundary extension of \( R_c \)
distance (This step is for the parallel version only)
1 create adjacency list \( G \) for all node in \( V_o \) using \( R_b \) as cutoff distance
2 for every vertex \( V \in V_o \)
   for each vertex pair \( V_i \) and \( V_j \) in \( \text{Neighbors}(V) \) do
     \( A_1 = \{V_i\} \)
     \( A_2 = \{V_j\} \)
     \( L_v = 0 \)
     while \( (A_1 \cap A_2 = \emptyset \text{ AND } L_v < L_{\text{max}}) \) do
       \( L_v = L_v + 2 \)
       if \( (A_1 \cap \text{Neighbors}(A_2) \neq \emptyset \text{ OR } A_1 \cap \text{Neighbors}(A_1) \neq \emptyset) \)
         \( L_v = L_v + 1 \)
       break
     else if \( (\text{Neighbors}(A_1) \cap \text{Neighbors}(A_2) \neq \emptyset) \)
       \( L_v = L_v + 2 \)
       \( A_1 = \text{Neighbors}(A_1) \)
       \( A_2 = \text{Neighbors}(A_2) \)
     if \( (L_v < L_{\text{max}}) \) \( \text{ \( K_v(L_v) \) was sent to \( \text{Neighbors}(V) \)} \)
### Spatial Hash-Function Tagging

**Algorithm** spatial hash function tagging (SHAFT)

**Input:**
- \( C(V) = 3D \) coordinates of all vertices (i.e., atoms)
- \( R_c = \) Ring cutoff range (Euclidean)
- \( R_{bc} = \) Bond cutoff distance (Euclidean)
- \( L_{\text{MAX}} = \) Maximum length of ring (integer)

**Output:**
- The integer index that is unique for all vertices in the maximum ring span

\[
\begin{align*}
  b &= R_{\text{lower}} / \sqrt{3} \\
  c &= R_{\text{upper}} L_{\text{MAX}} \\
  m &= \left\lfloor c/b \right\rfloor
\end{align*}
\]

**Step:**
- for each vertex
  - for each spatial dimension \( i \) from 1 to 3
    - \( q_i = \left\lfloor C_i/b \right\rfloor \)
    - \( q_i \mod m = m \)
  - return \( q = q_3 \times m^2 + q_2 \times m + q_1 \)

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
<td>20</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
</tbody>
</table>
Numerical Tests

Linear scaling on the problem size

Superlinear (strong) scaling on the number of CPUs
Dislocation Mining

Based on potential energy
Shown atoms with high energy compared to bulk

Based on shortest-path ring statistics
Shown atoms with less than 12 6-membered rings
100 km/s Impact on Notched AlN

- Dislocation nucleation & emission from notch during impact
- Dislocations & surface atoms mined by ring statistics
Impact-Damage Tolerant Ceramics?

Inverse problem: design materials with desired properties

209 million atom MD of hypervelocity impact in AlN for the design of light-weight ceramic armors
Crack Nucleation at Kink Bands

- Series of dislocation dipoles with opposite Burgers vectors form a kink band to release stress.
- Tilt grain boundaries of the kink bands act as sources of mode-II (shear) crack nucleation.
Dislocation Loops at Kink Bands

Graph (shortest-path circuit) based mining of topological defects

Atoms participating in non-6-member circuits

Dislocation network
Nanoindentation on Nanophase SiC

Superhardness

MD: 39 GPa
(grain size, $d = 8$ nm)

Expt.: 30-50 GPa
($d = 5-20$ nm)

[Liao et al., APL, ’05]

Szlufarska, Nakano & Vashishta, Science 309, 911 (’05)

Crossover from intergrain continuous response to intragranular discrete response
Multimodal Multidisplay Visualization

Hypervelocity penetration of an AlN plate
Singular Value Decomposition & Data Mining

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Dept. of Computer Science, Dept. of Physics & Astronomy,
Dept. of Chemical Engineering & Materials Science,
Dept. of Biological Sciences
University of Southern California

Email: anakano@usc.edu

Data mining $\cong$ data compression
**Rank of a Matrix**

- \( N \times M \) matrix \( A \) as a mapping: \( R^M \rightarrow R^N \)

\[
\begin{bmatrix}
1 \\
\vdots \\
M
\end{bmatrix}
\begin{bmatrix}
x \\
\vdots \\
x
\end{bmatrix}
(\in R^M)
\xrightarrow{A}
\begin{bmatrix}
b \\
\vdots \\
b
\end{bmatrix}
(\in R^N)
\]

- **Range of \( A \):** Vector space \( \{ b = Ax | \forall x \} \)

- **Rank of \( A \):** Number of linearly-independent vectors in the range, *i.e.*, how many linearly-independent \( N \)-element vectors are there in the range, such that

\[
b = A \forall x = \sum_{\nu=1}^{m} c_\nu | \nu \rangle
\]
Low Rank Approximations of a Matrix

- **Rank-1 approximation:** $NM \rightarrow N + M$
  \[
  \begin{bmatrix}
  M \\
  \psi \\
  \end{bmatrix}
  \begin{bmatrix}
  N \\
  \end{bmatrix}
  \approx
  \begin{bmatrix}
  u \\
  v \\
  \end{bmatrix}
  \]

- **Rank-2 approximation:** $NM \rightarrow 2(N + M)$
  \[
  \begin{bmatrix}
  \psi \\
  \end{bmatrix}
  \approx
  \begin{bmatrix}
  u_1 w_1[v_1] \\
  u_2 w_2[v_2] \\
  \end{bmatrix}
  \]

- **Rank-$m$ ($m \ll N, M$) approximation:** $NM \rightarrow m(N + M)$
  \[
  \begin{bmatrix}
  \psi \\
  \end{bmatrix}
  \approx
  \sum_{\nu=1}^{m}
  \begin{bmatrix}
  u_{\nu} \\
  \end{bmatrix}
  w_{\nu}[v_{\nu}]
  \]
Singular Value Decomposition

- **Problem:** Optimal approximation of an $N \times M$ matrix $\psi$ of rank-$m$ ($m \ll N$)?
- **Theorem:** An $N \times M$ matrix $\psi$ (assume $N \geq M$) can be decomposed as

$$
\psi = UDV^T = \sum_{\nu=1}^{M} U_{i\nu} d_\nu V_{j\nu} = \sum_{\nu=1}^{M} u_{i(\nu)} d_\nu v_{j(\nu)}
$$

where $U \in \mathbb{R}^{N \times M}$ & $V \in \mathbb{R}^{M \times M}$ are column orthogonal & $D$ is diagonal

$$
U^T U = V^T V = I_M
$$

- **Theorem:** Sort the SVD diagonal elements in descending order $d_1 \geq d_2 \geq ...$ & retain the first $m$ terms

$$
\psi^{(m)} = \sum_{\nu=1}^{m} u_{i(\nu)} d_\nu v_{j(\nu)}T
$$

which is optimal among $\forall$ rank-$m$ matrices in the 2-norm sense with the error

$$
\min_{\text{rank}(A)=m} \|A - \psi\|_2 = \|\psi^{(m)} - \psi\|_2 = d_{m+1}
$$
SVD for Image Compression

Original Image  5 Iterations  10 Iterations

20 Iterations  60 Iterations  100 Iterations

D. Richards & A. Abrahamsen
SVD in Data Mining

Given Point Set

Approximating Attributes by Representative Vectors

N. Ramakrishnan & A. Y. Grama
Machine Learning in Simulation

Aiichiro Nakano

Collaboratory for Advanced Computing & Simulations
Dept. of Computer Science, Dept. of Physics & Astronomy,
Dept. of Chemical Engineering & Materials Science,
Dept. of Biological Sciences
University of Southern California

Email: anakano@usc.edu
Rapid Genome Sequencing

- $10M Archon X prize for decoding 100 human genomes in 10 days & < $10K per genome (http://genomics.xprize.org): Preemptive attack on diseases

- Quantum tunneling current for rapid DNA sequencing?

Tsutsui et al., *Nature Nanotechnology* 5, 286 ('10);
Girdhar et al., *PNAS* 110, 16748 ('13)

- Tunneling current alone cannot distinguish the 4 nucleotides (A, C, G, T)
Rapid DNA Sequencing via Data Mining

- Use tunneling current (I)-voltage (V) characteristic (or electronic density-of-states) as the ‘fingerprints’ of the 4 nucleotides

![Tunneling current (I)-voltage (V) characteristic](image)

- Principal component analysis (PCA) & fuzzy c-means clustering clearly distinguish the 4 nucleotides

![Principal component analysis and fuzzy c-means](image)

- Viterbi algorithm for even higher-accuracy sequencing

H. Yuen et al, IJCS 4, 352 ('10)