**Novel Task Definition**

- **Motivation** – current search tools are insufficient
  - Prior knowledge of document contents is required to construct a query
  - Structure indicating relationships among returned documents is missing

- **Proposed Task** – select high-quality set of diverse threads in data graph
  - Node size indicates quality, edge length indicates node dissimilarity
  - Goal: select compact, high-quality paths that are well-separated

- **Example** – data elements are nodes
  - Related threading work
    - Selecting a single thread (D. Shahaf and C. Guastri, KDD 2010)
    - Constructing diverse topic threads (A. Ahmed and E. Xing, UAI 2010)

**Approach: Structured Determinantal Point Processes**

- Decompose thread quality and similarity
  - Score a set of threads \( Y \) via structured determinantal point process (SDPP)
  - SDPP: defines a distribution over sets \( Y \)
  - \( L_Y = \frac{\det(L_Y)}{\det(L + 1)} \)
  - \( P(Y) = \frac{\det(L_Y)}{\sum_{Y \subseteq \{1, \ldots, n\}} \det(L_Y)} \)
  - \( Y = \{i\} \rightarrow P(Y) \propto q(y_i)^2 \)
  - \( Y = \{i, j\} \rightarrow P(Y) \propto q(y_i)^2 q(y_j)^2 (1 - (\phi(y_i)^T \phi(y_j))^2) \)

- k-SDPPs: fix \( k \) # of points in \( Y \) to \( k \) (A. Kulesza and B. Taskar, ICML 2011)

- Sampling from k-SDPPs can be done in \( O(TmD^2 + D^3) \)

**Random Projection for Tractability**

- Complexity \( D^3 \) can be prohibitively large, so we project \( D \) down to \( d \)

- **Theorem** – Given \( \tilde{P}^k(Y) \) distribution after projecting \( D \) to \( d = O(\max(1/k, (\log(1/\epsilon) + \log N)/\epsilon^2)) \), error is bounded by:
  \[
  \| P^k - \tilde{P}^k \|_1 \leq e^{6\epsilon k} - 1 \approx 6\epsilon k
  \]
  with probability at least \( 1 - \delta \)

**Geographical Paths**

- **Data** – Cora, a large collection of computer science papers
- **Graph** – edges are citations
- **Figure** – example threads from a 4-SDPP with thread length \( T = 5 \); beside each thread are a few of its maximum-tfidf words
- **Baselines** – k-means clustering on time slices, dynamic topic model (DTM) (D. Blei and J. Lafferty, ICML 2006)

**How Det Balances Diversity and Quality**

- \( \det(x_1, x_2) \)
  - \( \begin{align*}
    \det(x_1, x_2) &= 3 & 1 \\
    1 & 2 & 4 \\
    2 & 2 & 4 \\
    1 & 3 & 4 \\
    3 & 1 & 8 \\
    1 & 3 & 8
  \end{align*} \)

- \( \det(x_2, x_3) \)
  - \( \begin{align*}
    \det(x_2, x_3) &= 5 & 5 \\
    1 & 3 & 10 \\
    5 & 5 & 10 \\
    1 & 3 & 10 \\
    3 & 1 & 8 \\
    1 & 3 & 8
  \end{align*} \)

**New York Times Timelines**

- **Data** – six 6-month NYT article sets
- **Graph** – edges are tfid cosine scores
- **Baselines** – k-means clustering on time slices, dynamic topic model (DTM)

<table>
<thead>
<tr>
<th>Intra-sim</th>
<th>Inter-sim</th>
<th>Human-sim</th>
<th>Precision/Recall</th>
<th>Vocab Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>8.28</td>
<td>2.01</td>
<td>4.32</td>
<td>11.23</td>
</tr>
<tr>
<td>DTM</td>
<td>14.47</td>
<td>0.71</td>
<td>3.78</td>
<td>8.06</td>
</tr>
<tr>
<td>k-SDPP</td>
<td>21.21</td>
<td>7.99</td>
<td>8.26</td>
<td>14.42</td>
</tr>
<tr>
<td>2D</td>
<td>8.06 / 2.18</td>
<td>19.443</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table: Intrinsic evaluation.** Intra-sim: Within-thread similarity (lower is better). Inter-sim: Between-thread similarity (lower is better). Human summary comparison. Human-sim: Cosine similarity. Precision: For each of the 10% highest-tfidf words in a filtered corpus, precision is \# words found in both divided by \# found in the threads.