Large-Scale Collection Threading using Structured k-DPPs

Alex Kulesza, Jennifer Gillenwater, Ben Taskar
University of Pennsylvania

Novel Task Definition

- **Motivation** – current search tools are insufficient

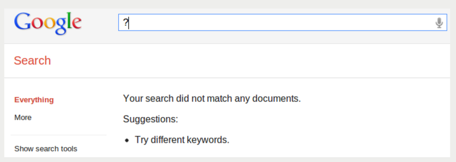
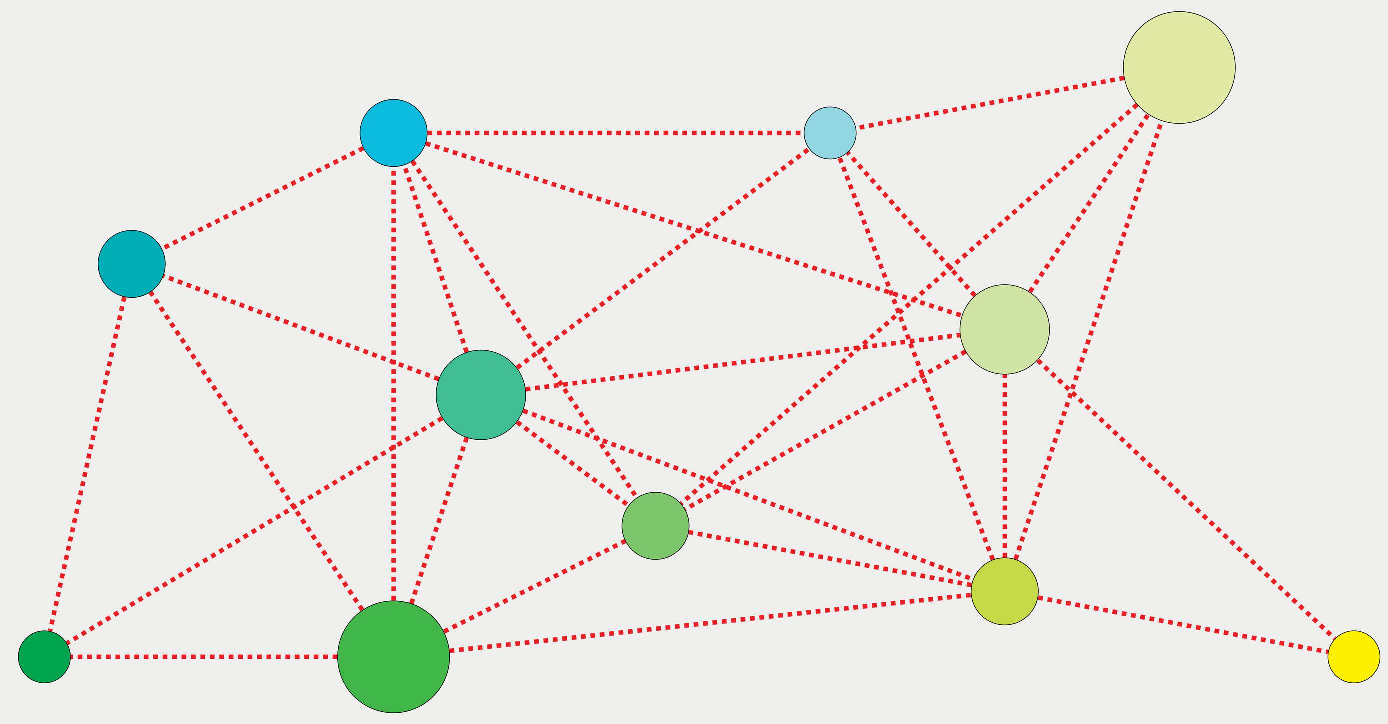


Figure: Prior knowledge of document contents is required to construct a query

Figure: Structure indicating relationships among returned documents is missing

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



- Related threading work
  - Selecting a *single* thread (D. Shahaf and C. Guestrin, KDD 2010)
  - Constructing diverse *topic* threads (A. Ahmed and E. Xing, UAI 2010)

Approach: Structured Determinantal Point Processes

- Decompose thread quality and similarity

$$q(y_i) = \prod_{t=1}^{T} q(y_{it}) \qquad \phi(y_i) = \sum_{t=1}^{T} \phi(y_{it})$$

- Score a set of threads Y via structured determinantal point process (SDPP) (A. Kulesza and B. Taskar, NIPS 2010)
- SDPP: defines a distribution over sets Y

$$L_{ij} = q(y_i)\phi(y_i)^\top \phi(y_j) q(y_j)$$

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\sum_{Y' \subseteq \{1,\dots,n\}} \det(L_{Y'})} = \frac{\det(L_Y)}{\det(L+I)}$$

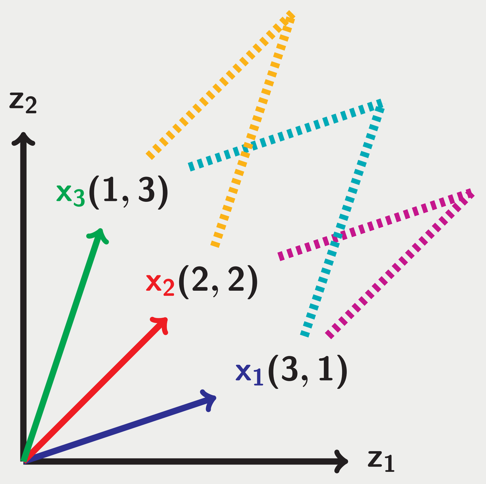
$$Y = \{i\} \rightarrow \mathcal{P}(Y) \propto q(y_i)^2$$

$$Y = \{i,j\} \rightarrow \mathcal{P}(Y) \propto q(y_i)^2 q(y_j)^2 (1 - (\phi(y_i)^\top \phi(y_j))^2)$$

- **k**-SDPPs: fix # of points in Y to k (A. Kulesza and B. Taskar, ICML 2011)
- Sampling from **k**-SDPPs can be done in $O(TrnD^2 + D^3)$

T = thread length $\qquad r$ = maximum node degree
n = # of nodes $\qquad D$ = # of features

How Det Balances Diversity and Quality

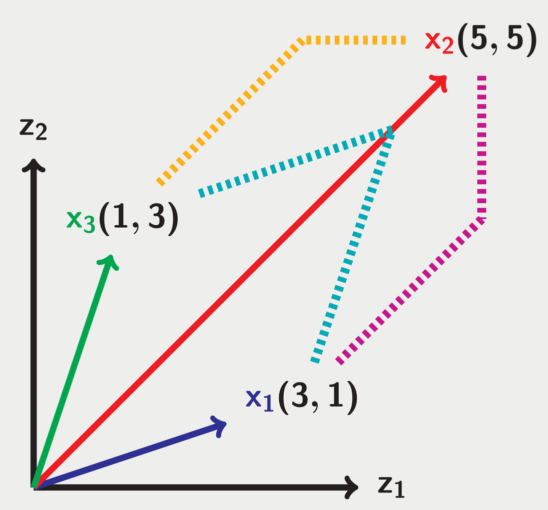


Diversity

$$\det(x_1, x_2) = \begin{vmatrix} 3 & 1 \\ 2 & 2 \end{vmatrix} = 4$$

$$\det(x_2, x_3) = \begin{vmatrix} 2 & 2 \\ 1 & 3 \end{vmatrix} = 4$$

$$\det(x_1, x_3) = \begin{vmatrix} 3 & 1 \\ 1 & 3 \end{vmatrix} = 8$$



Quality

$$\det(x_1, x_2) = \begin{vmatrix} 3 & 1 \\ 5 & 5 \end{vmatrix} = 10$$

$$\det(x_2, x_3) = \begin{vmatrix} 5 & 5 \\ 1 & 3 \end{vmatrix} = 10$$

$$\det(x_1, x_3) = \begin{vmatrix} 3 & 1 \\ 1 & 3 \end{vmatrix} = 8$$

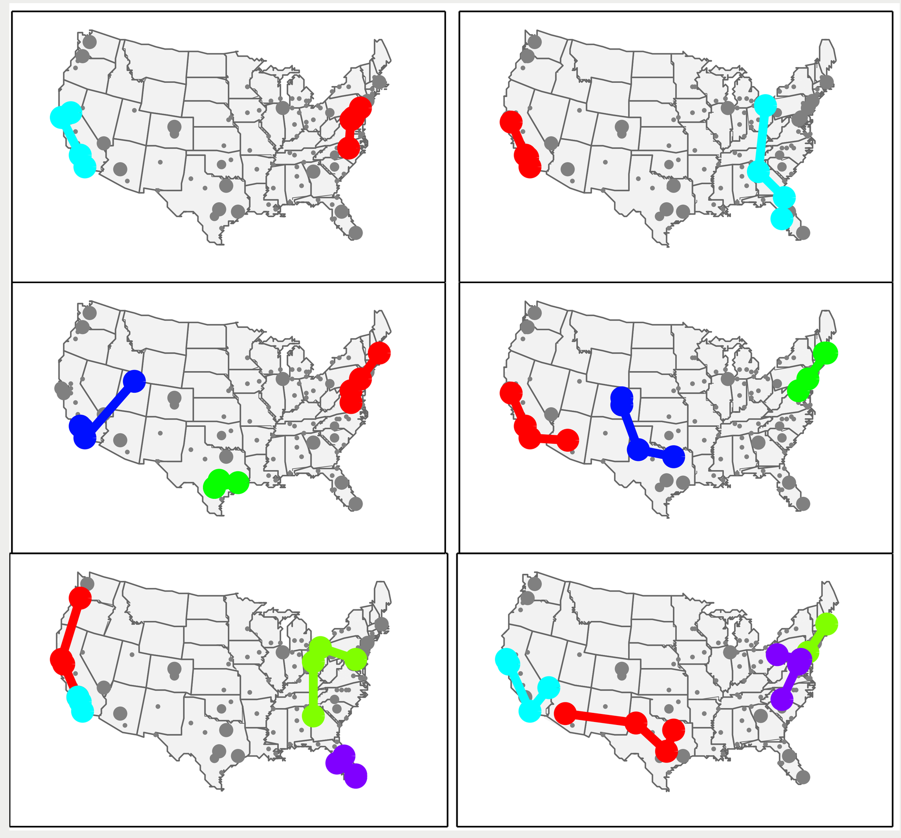
Random Projection for Tractability

- 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 $d = O(\max\{k/\epsilon, (\log(1/\delta) + \log N)/\epsilon^2\})$, error is bounded by:

$$\|P^k - \tilde{P}^k\|_1 \leq e^{6k\epsilon} - 1 \approx 6k\epsilon$$

with probability at least $1 - \delta$

Geographical Paths



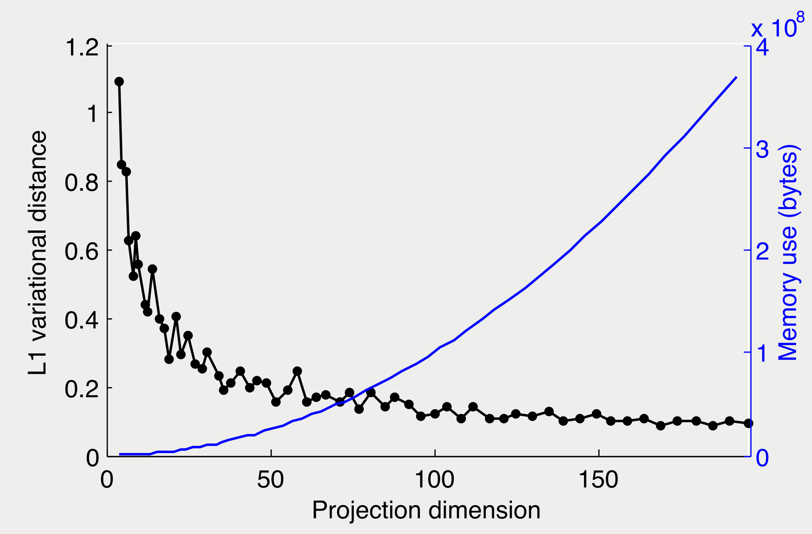
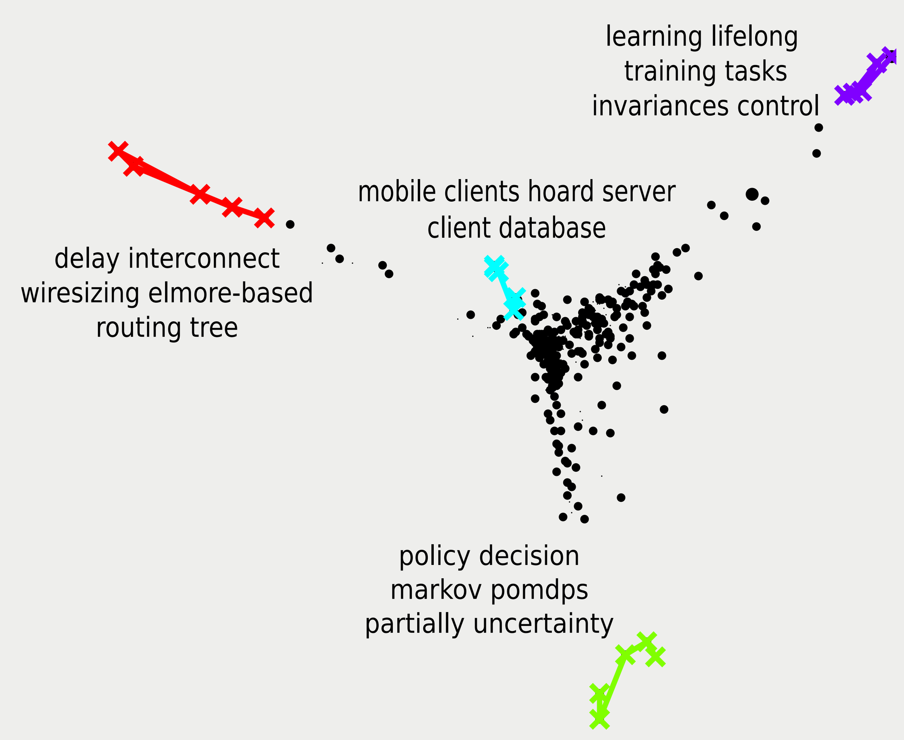


Figure: **Left**: Each row shows samples drawn from a **k**-SDPP with path length $T = 4$. City size indicates Google hit count. From top to bottom, $k = 2, 3, 4$. **Above**: Effects of random projections.

Cora Citation Threads

- **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 (we project from word-space to 2D via PCA on thread centroids)

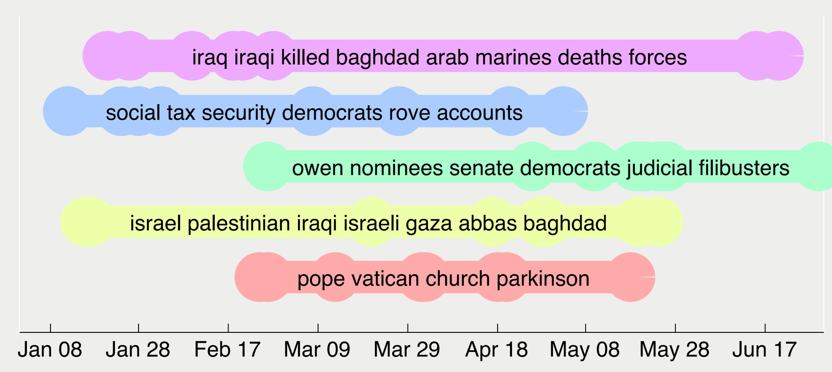


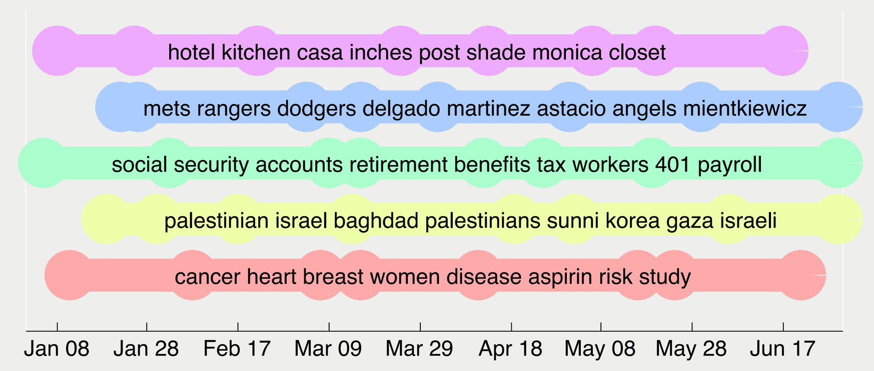
New York Times Timelines

- **Data** – six 6-month NYT article sets; **Graph** – edges are tfidf cosine scores
- **Baselines** – **k**-means clustering on time slices, dynamic topic model (DTM) (D. Blei and J. Lafferty, ICML 2006)

	Intra-sim	Inter-sim	Human-sim	Precision/Recall	Time (sec)
k-means	8.28	2.01	4.32	11.23 / **7.28**	625
DTM	14.47	**0.71**	3.78	8.06 / 2.18	19,443
k-SDPP	**21.21**	7.79	**8.26**	**14.42** / 5.86	**252**

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





Feb 24: Parkinson's Disease Increases Risks to Pope
Feb 26: Pope's Health Raises Questions About His Ability to Lead
Mar 13: Pope Returns Home After 18 Days at Hospital
Apr 01: Pope's Condition Worsens as World Prepares for End of Papacy
Apr 02: Pope, Though Gravely Ill, Utters Thanks for Prayers
Apr 18: Europeans Fast Falling Away from Church
Apr 20: In Developing World, Choice [of Pope] Met with Skepticism
May 18: Pope Sends Message with Choice of Name

Jan 11: Study Backs Meat, Colon Tumor Link
Feb 07: Patients—and Many Doctors—Still Don't Know How Often Women Get Heart Disease
Mar 07: Aspirin Therapy Benefits Women, but Not in the Way It Aids Men
Mar 16: Study Shows Radiation Therapy Doesn't Increase Heart Disease Risk for Breast Cancer Patients
Apr 11: Personal Health: Women Struggle for Parity of the Heart
May 16: Black Women More Likely to Die from Breast Cancer
May 24: Studies Bolster Diet, Exercise for Breast Cancer Patients
Jun 21: Another Reason Fish is Good for You

Figure: A set of five news threads sampled from a **k**-SDPP (left) and threads generated by a dynamic topic model (right). Above, threads are shown with the most salient words superimposed; below, headlines from the last thread are listed.