Large-Scale Modeling of Diverse Paths using k-SDPPs

Jennifer Gillenwater, Alex Kulesza, Ben Taskar
University of Pennsylvania

Motivation for document collection modeling

- Document collections are too big for manual examination

Figure: News articles

Figure: Research papers

- Current search tools are lacking

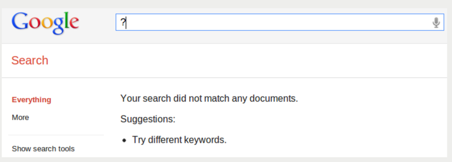


Figure: Prior knowledge of document contents is required to construct a query

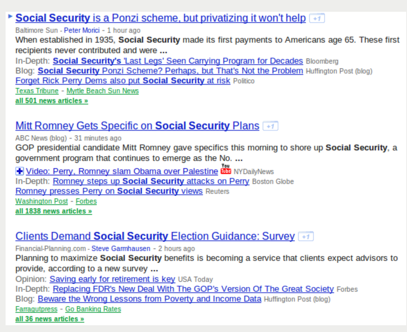
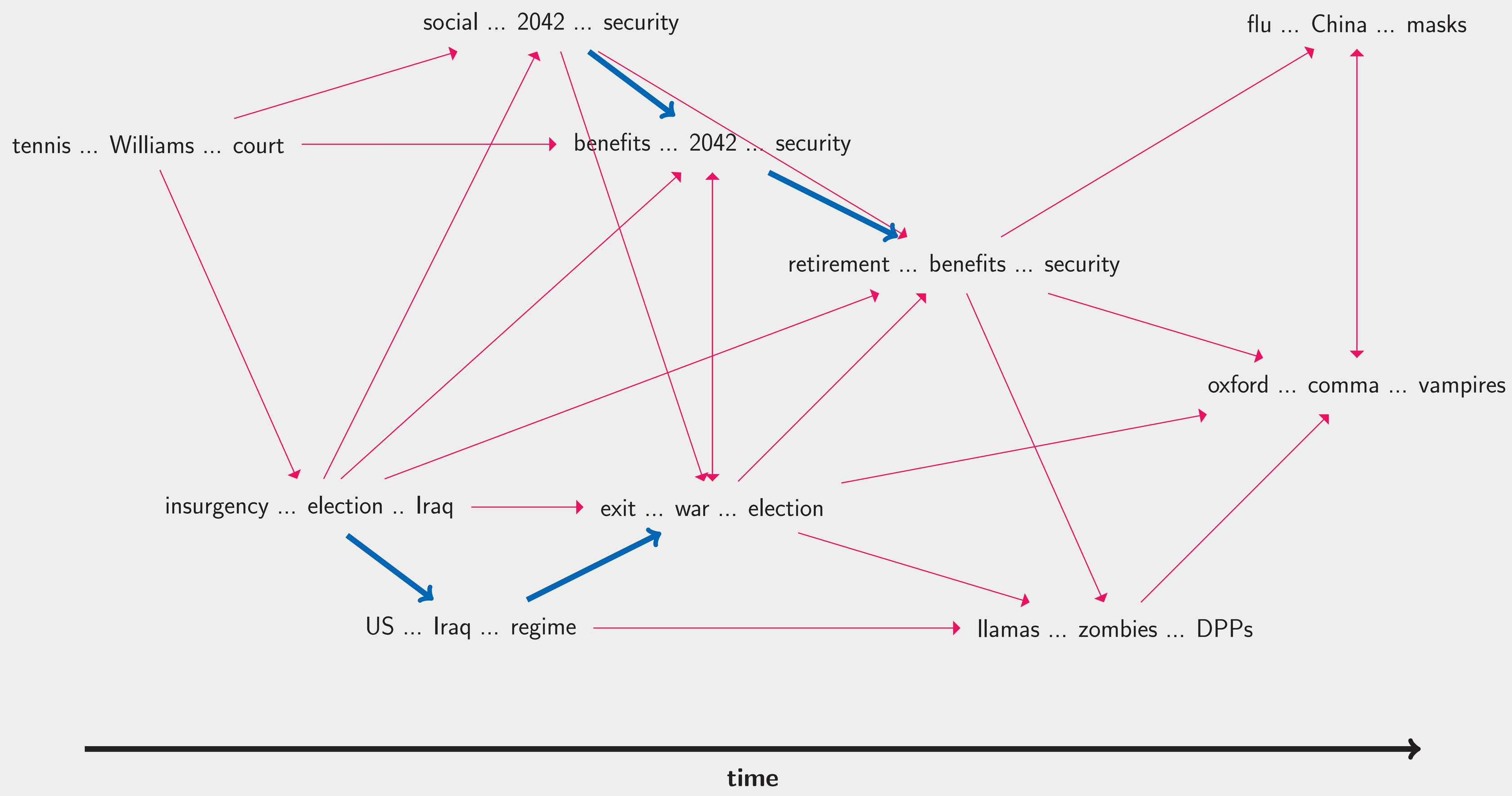


Figure: Structure indicating relationships among returned documents is missing

Novel problem definition

Select a high-quality set of diverse threads in a document graph.

- Graph nodes = documents
- Graph edges = document similarities (e.g. tfidf cosine scores)
- Thread = a path through the graph



Related document threading work

- Topic detection and tracking (TDT) program
- Selecting a *single* thread (D. Shahaf and C. Guestrin, KDD 2010)
- Constructing diverse *topic* threads (A. Ahmed and E. Xing, UAI 2010)

Our approach: determinantal point processes

- Decompose thread quality as a product over nodes $q(\mathbf{y}_i) = \prod_{t=1}^{T} q(y_{it})$
- Decompose thread similarity as a sum over nodes $\phi(\mathbf{y}_i) = \sum_{t=1}^{T} \phi(y_{it})$
- Score a set of threads $\mathbf{Y}$ using a determinantal point process (DPP)
- DPP: defines a distribution over sets $\mathbf{Y}$

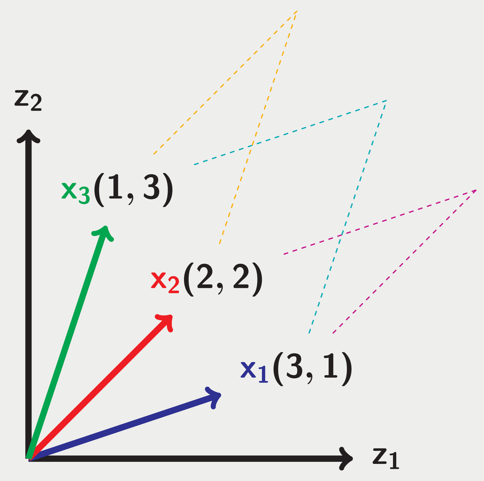
$$L_{ij} = q(\mathbf{y}_i)\phi(\mathbf{y}_i)^\top \phi(\mathbf{y}_j) q(\mathbf{y}_j)$$

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

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

$$\mathbf{Y} = \{i,j\} \rightarrow \mathcal{P}(\mathbf{Y}) \propto q(\mathbf{y}_i)^2 q(\mathbf{y}_j)^2 (1 - (\phi(\mathbf{y}_i)^\top \phi(\mathbf{y}_j))^2)$$

How determinants balance diversity and quality

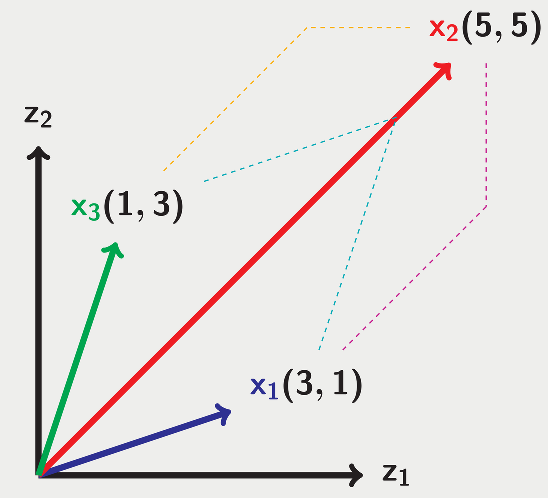


Diversity

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

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

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



Quality

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

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

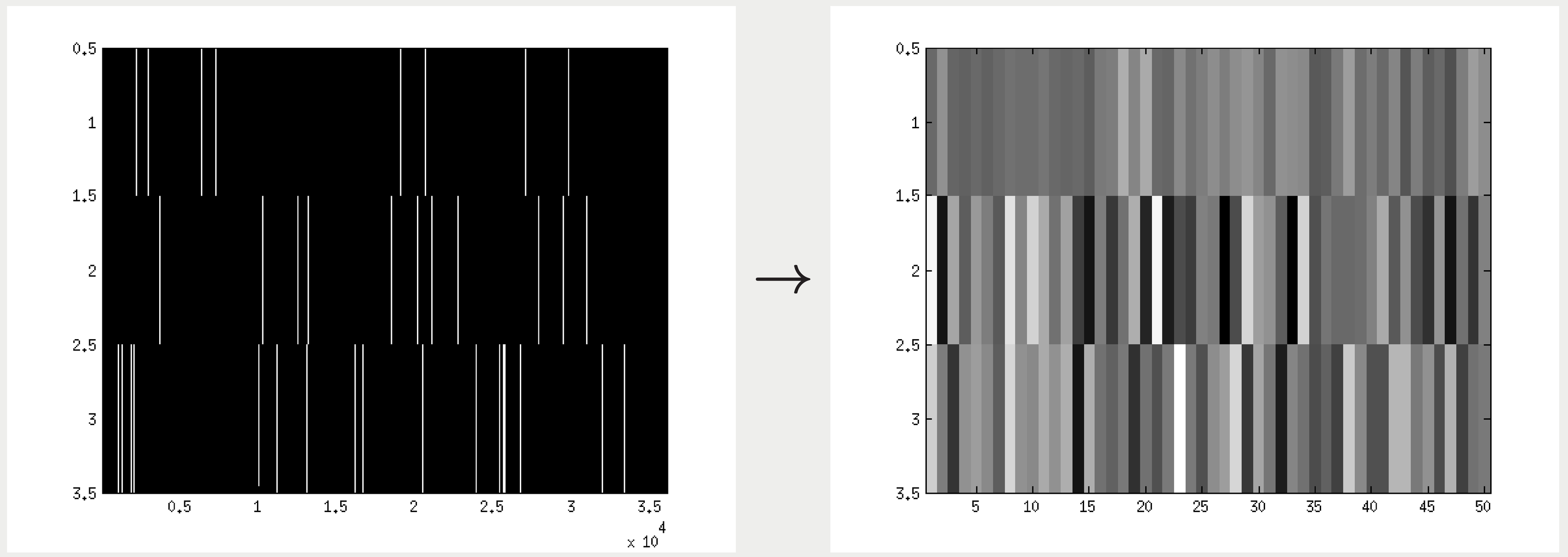
$$\det(x_1, x_3) = \begin{vmatrix} 3 & 1 \\ 1 & 3 \end{vmatrix} = 8$$

Random projection for tractability

- **k**-DPPs [4]: fix # of points in $\mathbf{Y}$ to **k**, s.t. $\mathcal{P}^k(\mathbf{Y}) = \frac{\det(L_{\mathbf{Y}})}{\sum_{|\mathbf{Y}'|=k} \det(L_{\mathbf{Y}'})}$
- Sampling from **k**-SDPPs can be done in $O(TrnD^2)$

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

- If there is one feature per word, $D > 30,000$
- With $n > 30,000$ also, nD^2 is prohibitively large
- Single message in sampling algorithm would be **200** terabytes
- **Theorem**: Given $\tilde{P}^k(\mathbf{Y})$ = 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 $1 - \delta$
- In practice, we projected D down to $d = 50$



Experiments on the New York Times

- Constructed graphs on 6-month periods of news articles
- Baseline 1: Clustering - split articles into **T** time slices and apply **k**-means
- Baseline 2: Non-max suppression - iterative sampling of threads
- **k**-SDPP - global thread-set optimization

	2005a	2005b	2006a	2006b	2007a	2007b	2008a	2008b
CLS	3.53	3.85	3.76	3.62	3.47	3.32	3.70	3.00
NMX	3.87	3.89	4.59	5.12	3.73	3.49	4.58	3.59
k-SDPP	6.91*	5.49*	5.79*	8.52*	6.83*	4.37*	4.77	3.91

Table: a: January-June, b: July-December. Star (*) implies significant at 99% confidence. Scoring metric: Cosine similarity between threads and human-generated news summaries.

References

1. D. Shahaf and C. Guestrin. Connecting the Dots Between News Articles. KDD 2010.
2. A. Ahmed and E. Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. UAI 2010.
3. A. Kulesza and B. Taskar. Structured Determinantal Point Processes. NIPS 2010.
4. A. Kulesza and B. Taskar. k-DPPs: Fixed-Size Determinantal Point Processes. ICML 2011.