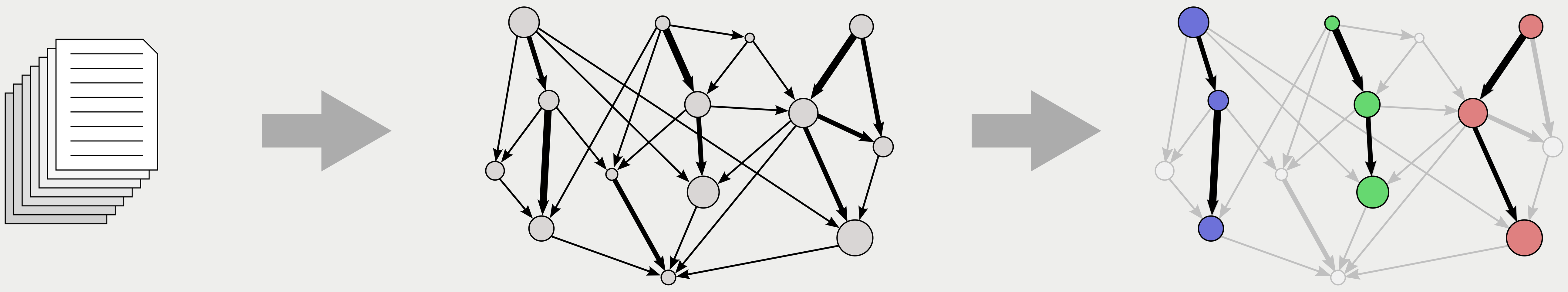
Discovering Diverse and Salient Threads in Document Collections

Jennifer Gillenwater, Alex Kulesza, Ben Taskar
University of Pennsylvania

Document Collection Threading – (1) Build a graph from the collection, using measures of importance and relatedness to weight nodes (documents) and build edges (relationships). (2) From this graph, extract a diverse, salient set of threads to represent the collection.



Introduction

- Motivation: current search tools are insufficient

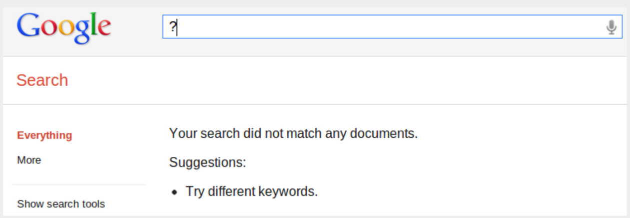


Figure: Prior knowledge of document contents is required to construct a query

Figure: Structure indicating relationships among returned documents is missing

- 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: Determinantal Point Processes

- Decompose quality and similarity of a thread $\mathbf{y_i} = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{iT}})$

$$q(\mathbf{y_i}) = q(\mathbf{y_{i1}}) \prod_{t=2}^{T} q(\mathbf{y_{it}}) q(\mathbf{y_{i(t-1)}}, \mathbf{y_{it}}) \qquad \phi(\mathbf{y_i}) = \sum_{t=1}^{T} \phi(\mathbf{y_{it}})$$

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

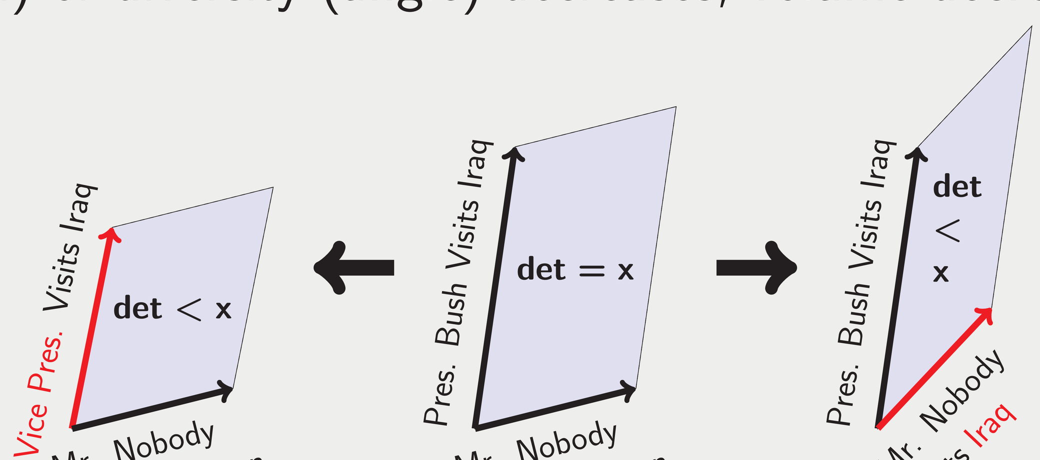
$$\mathbf{L_{ij}} = q(\mathbf{y_i})\phi(\mathbf{y_i})^\top \phi(\mathbf{y_j}) q(\mathbf{y_j})$$

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

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

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

- $\det(\mathbf{L_Y})$ is proportional to volume spanned by the vectors $q(\mathbf{y_i})\phi(\mathbf{y_i})$. As quality (length) or diversity (angle) decreases, volume decreases.



- $\mathbf{k}$-SDPPs: fix # of points in $\mathbf{Y}$ to $\mathbf{k}$ (A. Kulesza and B. Taskar, ICML 2011)
- Sampling from $\mathbf{k}$-SDPPs can be done in $\mathbf{O(TrnD^2 + D^3)}$, where $\mathbf{r}$ = max node degree, $\mathbf{n}$ = # of nodes, $\mathbf{D}$ = # of features

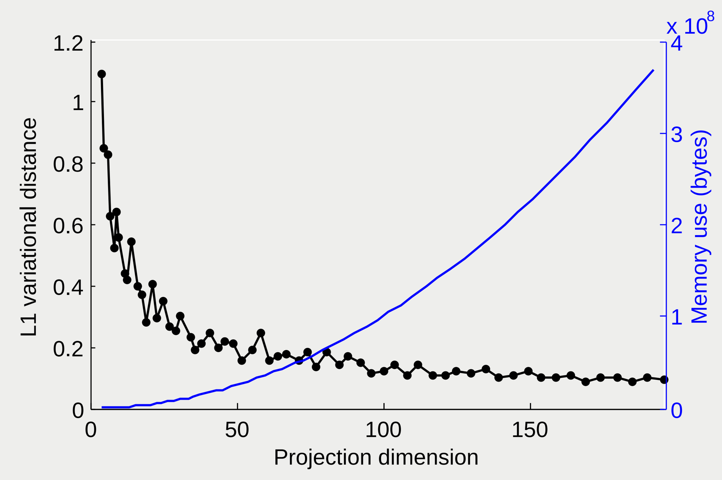
Random Projections for Tractability

- Complexity $\mathbf{D^3}$ can be prohibitively large, so we project $\mathbf{D}$ down to $\mathbf{d}$
- **Theorem**: Given $\tilde{\mathcal{P}}^{\mathbf{k}}(\mathbf{Y})$ = distribution after projecting $\mathbf{D}$ to $\mathbf{d} = \mathbf{O}(\max\{\mathbf{k}/\epsilon, (\log(1/\delta) + \log \mathbf{N})/\epsilon^2\})$, error is bounded by:

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

with probability at least $1 - \delta$.

- Random projections on a small threading task where the exact model is tractable: $\mathbf{n} = 600$ and $\mathbf{D} = 150$. As predicted by the theorem, fidelity to the true model increases rapidly with $\mathbf{d}$.

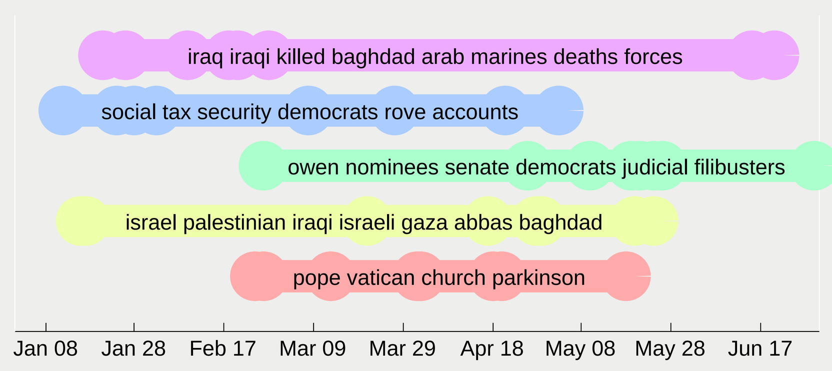


New York Times Timelines

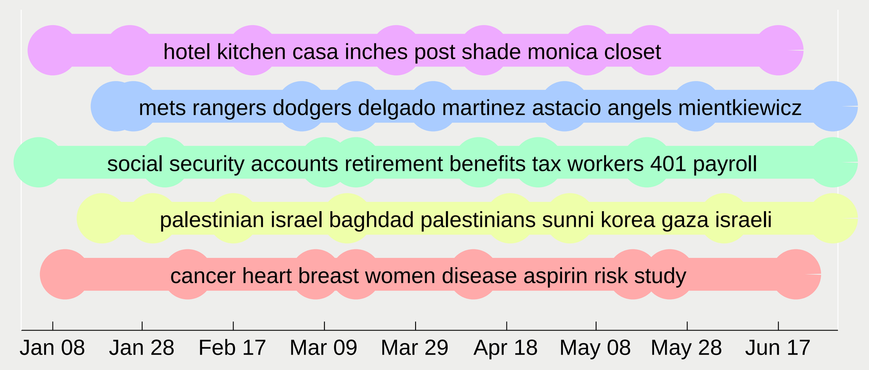
- **Data** – six 6-month NYT article sets; **Graph** – edges are tfidf cosine scores
- **Baselines** – **k**-means clustering on time slices, dynamic topic model (DTM) (D. Blei and J. Lafferty, ICML 2006)

	ROUGE-SU4	Coherence	Interlopers	Secs
k-means	3.76	2.73	0.71	625
DTM	3.44	3.19	1.10	19,443
k-SDPP	**3.98**	**3.31**	**1.15**	**252**

Table: **ROUGE-SU4**: comparison to human summaries. **Mechanical Turk**: thread coherence rating (1-5); average # of random interloper articles identified. **Secs**: runtime.



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 threads from a **k**-SDPP (left) and a DTM (right). Above, threads are shown with the most salient words superimposed; below, headlines from the last thread are listed.

Example New York Times Graph

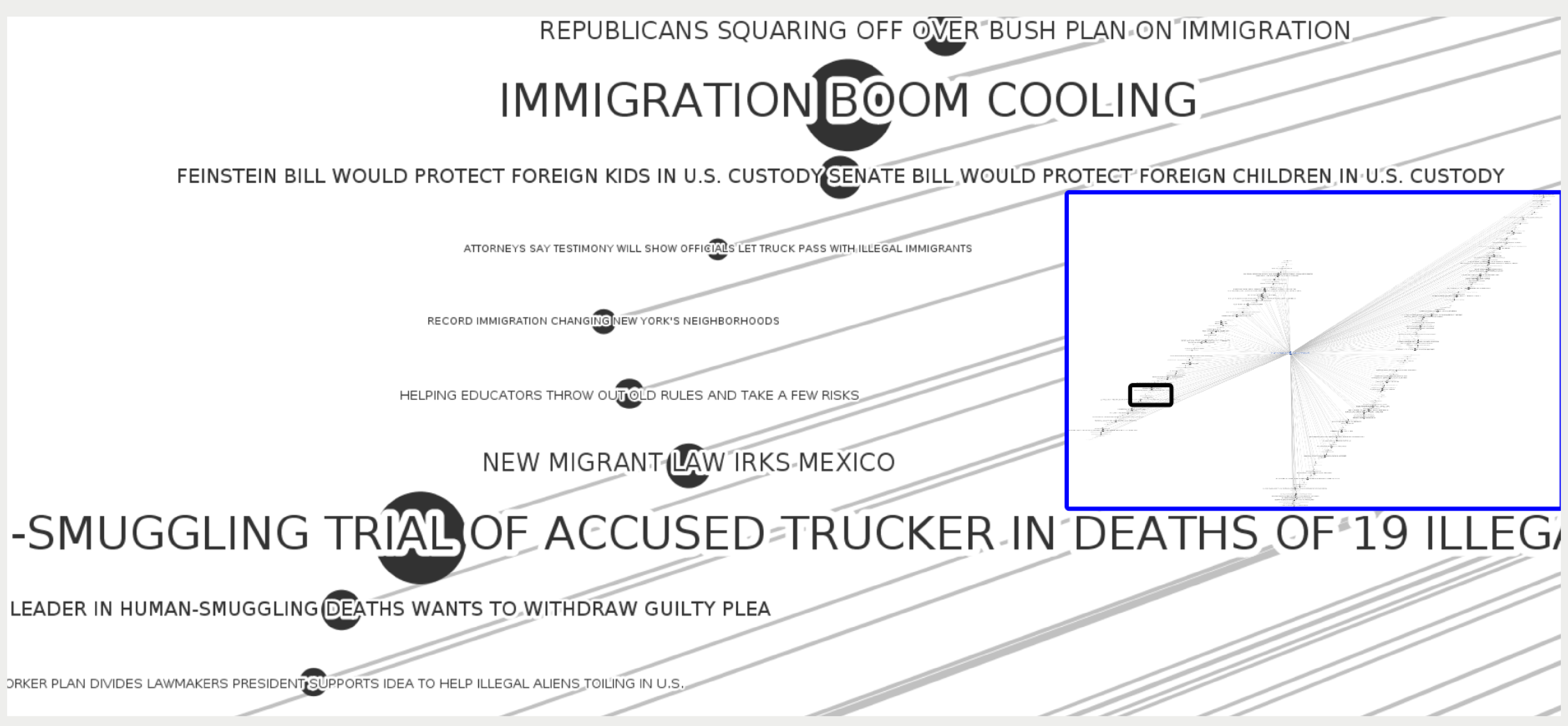


Figure: **Blue box**: Single node "Study Analyzes Data on Illegal Immigrants" with all its neighbors. **Other**: Zoom in, indicated by a black rectangle in the full image.

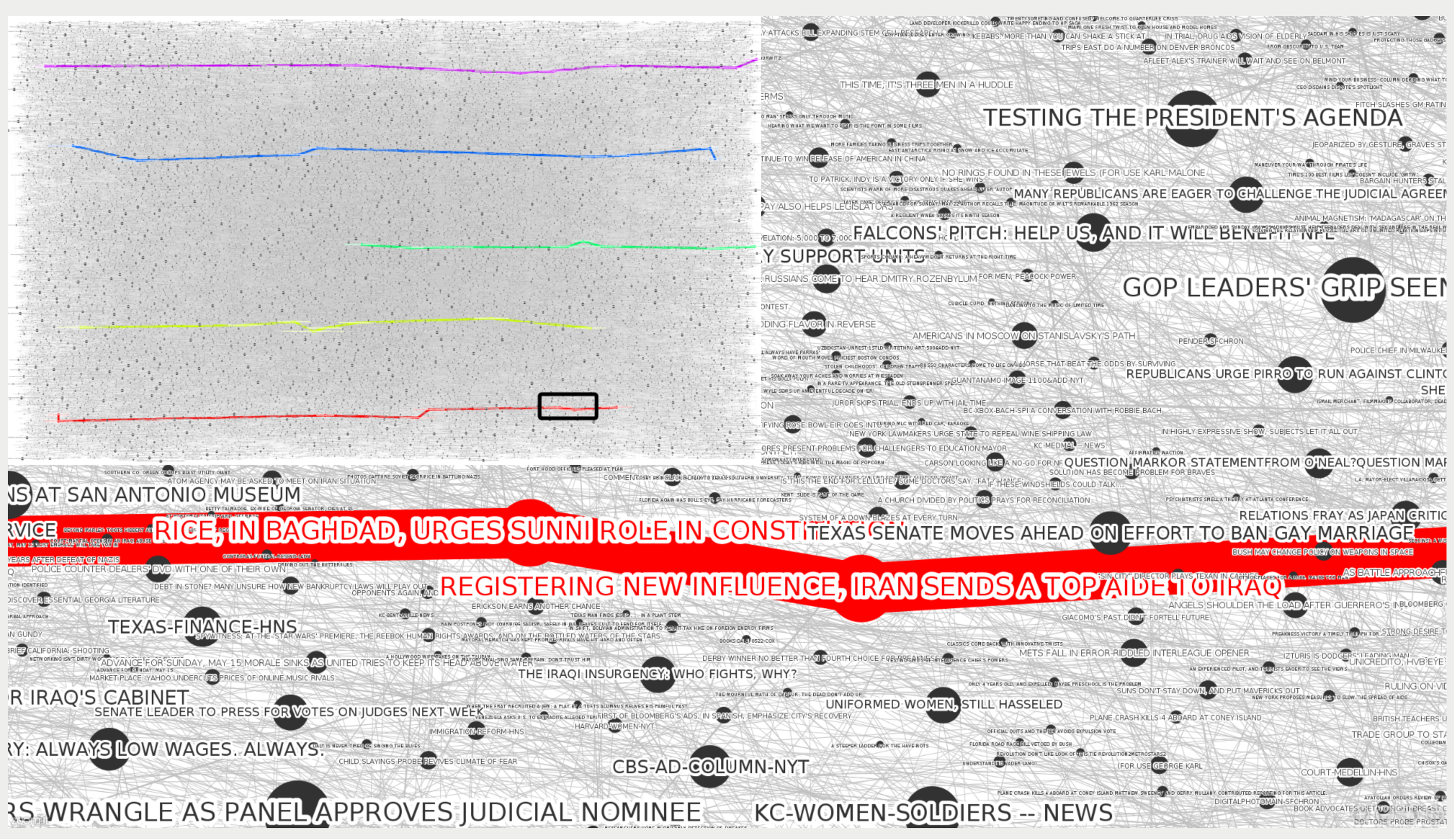


Figure: **Top left**: Full graph with 5 DPP threads. **Other**: Zoom in.