# 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

### Current search tools are lacking

Google	3	Ŷ
Search		
Everything More	Your search did not match any documents. Suggestions:	



Figure: Research papers

Social Security is a Ponzi scheme, but privatizing it won't help Balimore Sun - Peter Morid - 1 hour ago When established in 1935, Social Security made its first payments to Americans age 65. These first recipients never contributed and were ... In-Depth: Social Security's 'Last Legs' Seen Carrying Program for Decades Bloomberg Blog: Social Security Ponzi Scheme? Perhaps, but That's Not the Problem Huffington Post (blog) Forget Rick Perry Dems also put Social Security at risk Politico Texas Tribune - Myrtle Beach Sun News all 501 news articles =

 Mitt Romney Gets Specific on Social Security Plans
 +7

 ABC News (blog) - 31 minutes ago
 GOP presidential candidate Mitt Romney gave specifics this morning to shore up Social Security, government program that continues to emerge as the No. ...

 Video: Perry, Romney slam Obama over Palestine
 MYDalyNews
 In-Depth: Romney steps up Social Security attacks on Perry Boston Globe Romney presses Perry on Social Security views Reuters
 Washington Post - Earbes
 ItaBa new articles
 ItaBa new arthe set anem articles
 ItaBa new articles

# How determinants balance diversity and quality



Diversity







 $det(x_1, x_3) =$ 



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
```



# Random projection for tractability

▶ k-DPPs [4]: fix # of points in Y to k, s.t. P<sup>k</sup>(Y) = det(L<sub>Y</sub>)/∑|Y'|=k det(L<sub>Y'</sub>)
▶ Sampling from k-SDPPs can be done in O(TrnD<sup>2</sup>)
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<sup>2</sup> is prohibitively large
Single message in sampling algorithm would be 200 terabytes
Theorem: Given P<sup>k</sup>(Y) = distribution after projecting D to d = O(max{k/ε, (log(1/δ) + log N)/ε<sup>2</sup>}), error is bounded by: ||P<sup>k</sup> - P<sup>k</sup>||<sub>1</sub> ≤ e<sup>6kε</sup> - 1 ≈ 6kε
with probability at least 1 - δ

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

#### 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
S	3.53	3.85	3.76	3.62	3.47	3.32	3.70	3.00

Decompose thread similarity as a sum over nodes φ(y<sub>i</sub>) = ∑<sub>t=1</sub><sup>I</sup> φ(y<sub>it</sub>)
 Score a set of threads Y using a determinantal point process (DPP)
 DPP: defines a distribution over sets Y

$$\begin{split} \mathbf{L}_{ij} &= \mathbf{q}(\mathbf{y}_i) \boldsymbol{\phi}(\mathbf{y}_i)^\top \boldsymbol{\phi}(\mathbf{y}_j) \mathbf{q}(\mathbf{y}_j) \\ & \det(\mathbf{L}_{\mathbf{Y}}) \\ & \frac{\det(\mathbf{L}_{\mathbf{Y}})}{\sum_{\mathbf{Y}' \subseteq \{1,...,n\}} \det(\mathbf{L}_{\mathbf{Y}'})} = \frac{\det(\mathbf{L}_{\mathbf{Y}})}{\det(\mathbf{L}+\mathbf{I})} \end{split}$$

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

NMX3.873.894.595.123.733.494.583.59**k**-SDPP6.91\*5.49\*5.79\*8.52\*6.83\*4.37\*4.773.91Table: a: January-June, b: July-December. Star (\*) implies significant at 99% confidence.Scoring metric: Cosine similarity between threads and human-generated news summaries.

### References

CL

 D. Shahaf and C. Guestrin. Connecting the Dots Between News Articles. KDD 2010.
 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.
 A. Kulesza and B. Taskar. Structured Determinantal Point Processes. NIPS 2010.
 A. Kulesza and B. Taskar. k-DPPs: Fixed-Size Determinantal Point Processes. ICML 2011.