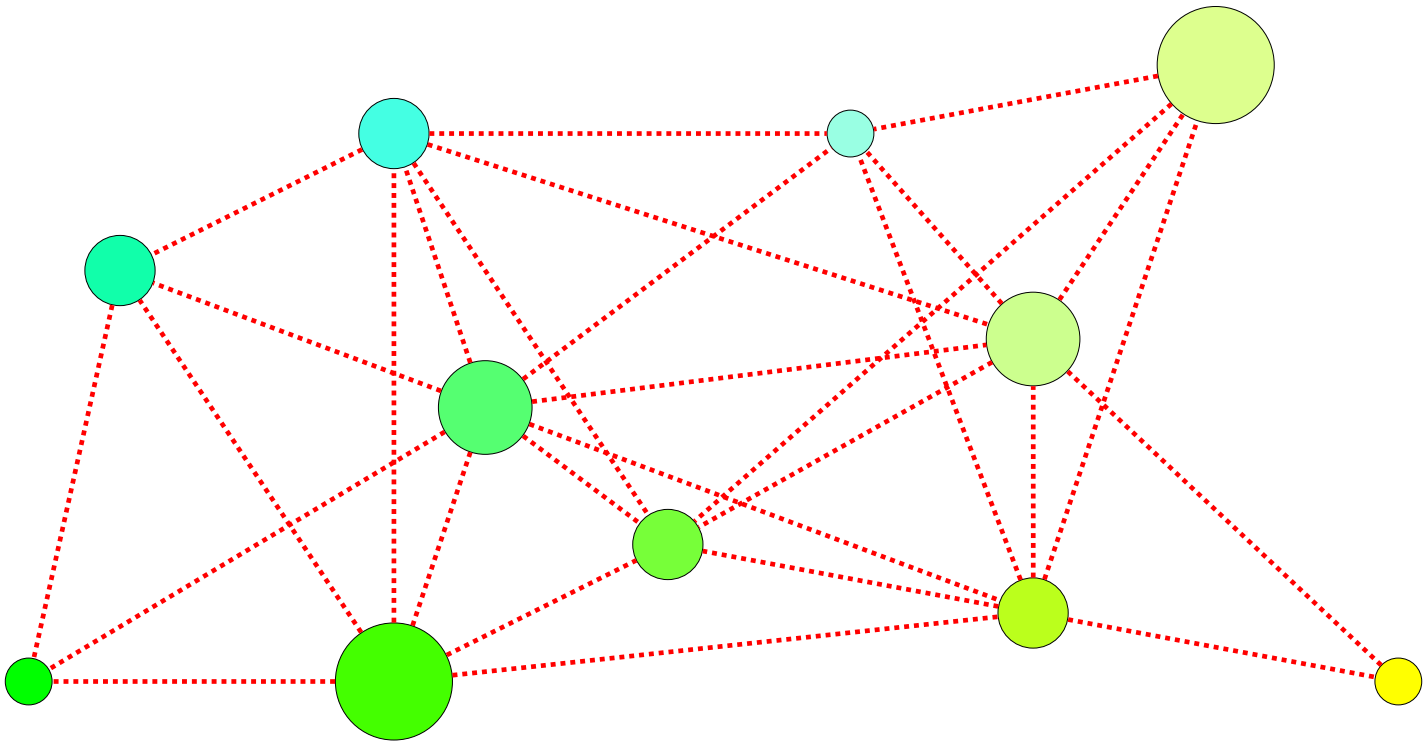
Large-Scale Collection Threading using Structured k-DPPs

Jennifer Gillenwater, Alex Kulesza, Ben Taskar

University of Pennsylvania

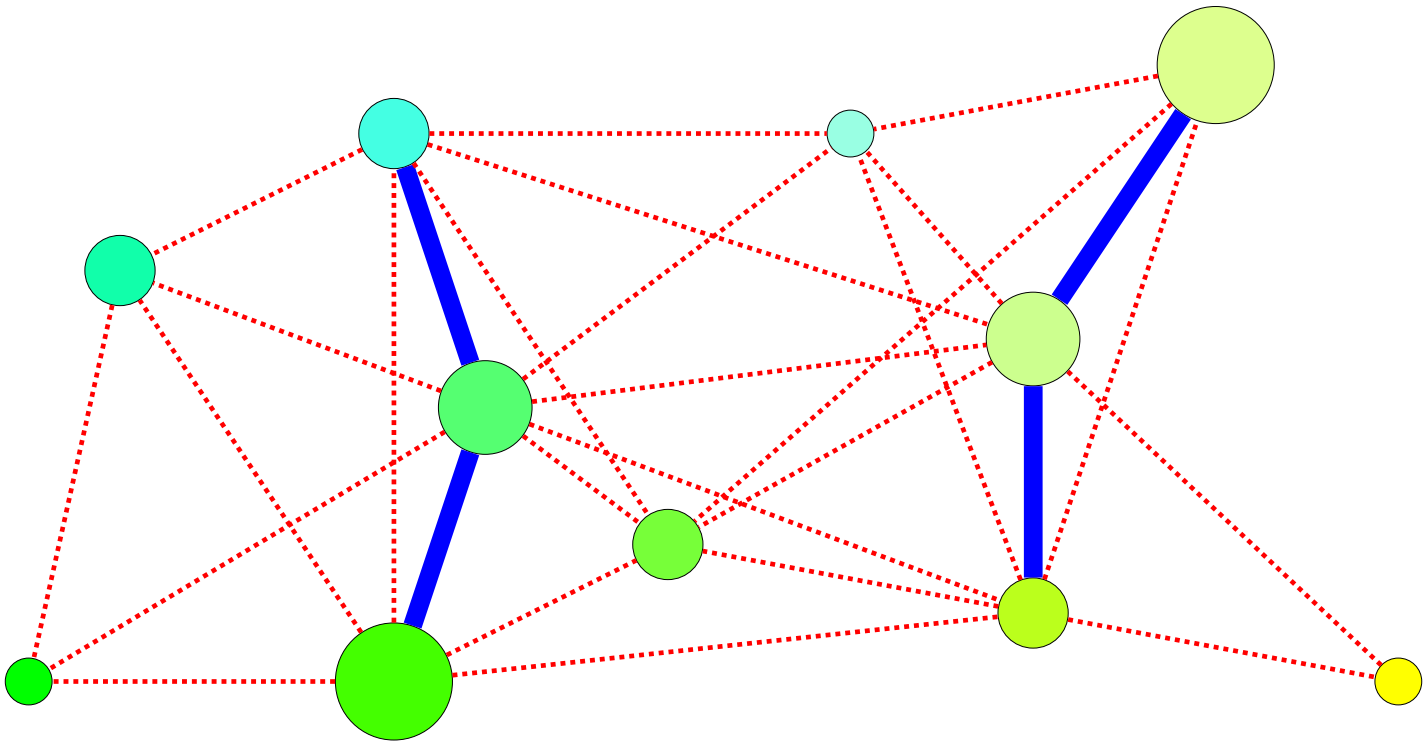
Novel task definition

Select a *high-quality set of diverse paths* in a data graph

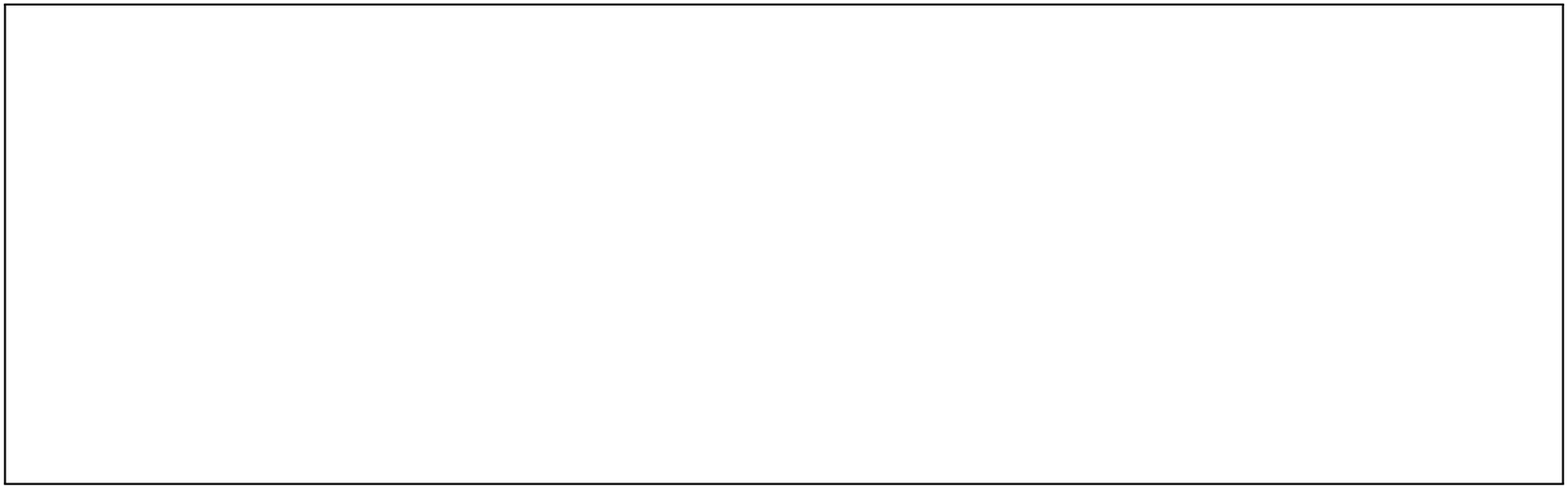


Novel task definition

Select a *high-quality set of diverse paths* in a data graph



News Timeline



News Timeline

Feb 24: Parkinson's Disease Increases Risks to Pope

News Timeline

Feb 24: Parkinson's Disease Increases Risks to Pope
Feb 26: Pope's Health Raises Questions About His Ability to Lead

News Timeline

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

News Timeline

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

News Timeline

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

News Timeline

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

News Timeline

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

News Timeline

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

Related threading work

- Selecting a *single* thread (D. Shahaf and C. Guestrin, KDD 2010)

Related threading work

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

Approach: Structured Determinantal Point Processes

Score a set of threads **Y** via structured determinantal point process (SDPP)[1]

[1] (A. Kulesza and B. Taskar, NIPS 2010)

Approach: Structured Determinantal Point Processes

Score a set of threads $\mathbf{Y}$ via structured determinantal point process (SDPP)[1]

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

[1](A. Kulesza and B. Taskar, NIPS 2010)

Approach: Structured Determinantal Point Processes

Score a set of threads $\mathbf{Y}$ via structured determinantal point process (SDPP)[1]

$$\mathcal{P}(\mathbf{Y}) \propto \det(\mathbf{L}_{\mathbf{Y}})$$

[1] (A. Kulesza and B. Taskar, NIPS 2010)

Approach: Structured Determinantal Point Processes

Score a set of threads $\mathbf{Y}$ via structured determinantal point process (SDPP)[1]

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

[1] (A. Kulesza and B. Taskar, NIPS 2010)

Approach: Structured Determinantal Point Processes

Score a set of threads $\mathbf{Y}$ via structured determinantal point process (SDPP)[1]

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

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

[1] (A. Kulesza and B. Taskar, NIPS 2010)

Approach: Structured Determinantal Point Processes

k-SDPPs[1]: fix # of points in **Y** to **k**

[1](A. Kulesza and B. Taskar, ICML 2011)

Sampling from **k**-SDPPs

$$\mathbf{O(TrnD^2 + D^3)}$$

$\mathbf{T}$ = thread length

$\mathbf{r}$ = maximum node degree

$\mathbf{n}$ = # of nodes

$\mathbf{D}$ = # of features

Random Projection for Tractability

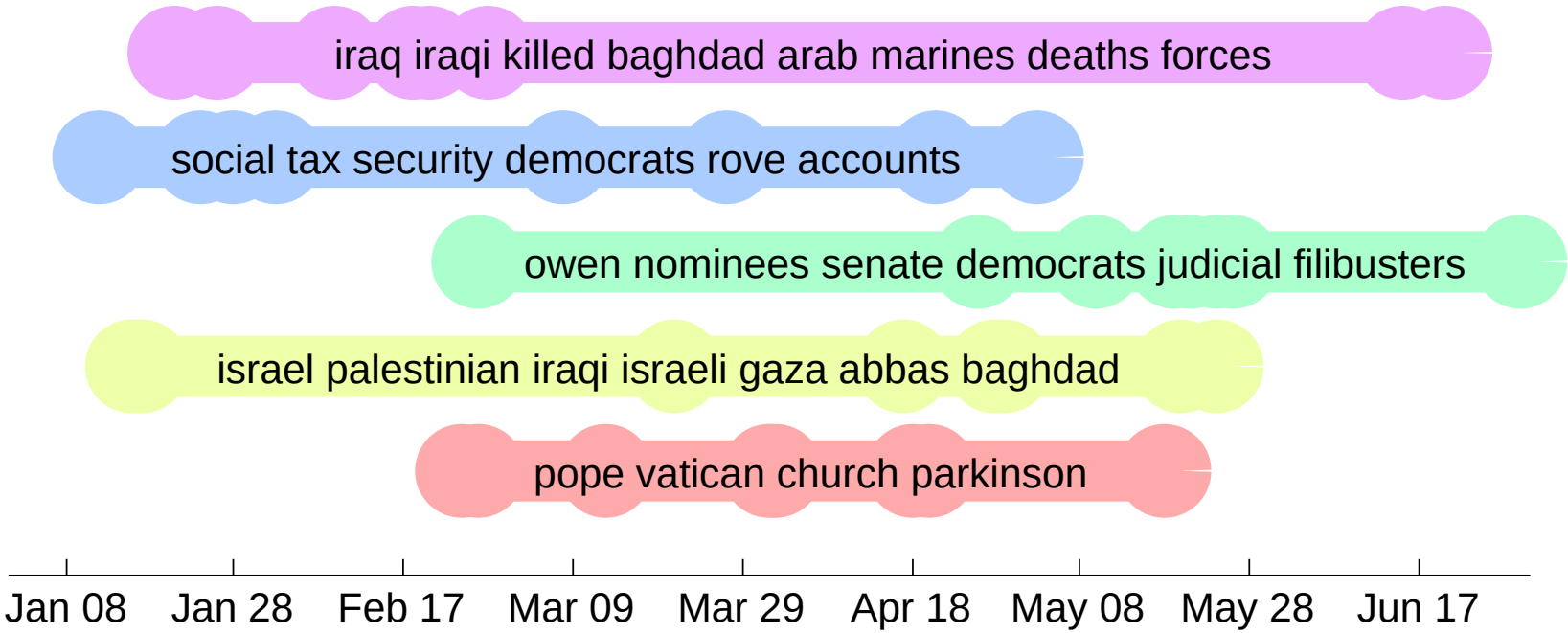
- Project **D** down to **d**

Random Projection for Tractability

- Project **D** down to **d**
- **Theorem**: Given $\tilde{\mathbf{P}}^{\mathbf{k}}(\mathbf{Y})$ = distribution after projecting

$$\|\mathbf{P}^{\mathbf{k}} - \tilde{\mathbf{P}}^{\mathbf{k}}\|_{\mathbf{1}} \leq \text{nice small things}$$

New York Times Timelines



New York Times Timelines

k-means DTM **k**-SDPP		

New York Times Timelines

	Human summary similarity	
k-means	4.32	
DTM	3.78	
k-SDPP	**8.26**	

New York Times Timelines

	Human summary similarity	Runtime (sec)
k-means	4.32	625
DTM	3.78	19,443
k-SDPP	**8.26**	**252**