# Large-Scale Modeling of Diverse Paths using Structured $k$-DPPs

Jennifer Gillenwater Alex Kulesza Ben Taskar
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
{jengi,kulesza,taskar}@cis.upenn.edu

December 16, 2012

Thanks to the increasing availability of large, interrelated document collections, we now have easy access to vast stores of information that are orders of magnitude too big for manual examination. Tools like search have made these collections useful for the average person; however, search tools require prior knowledge of likely document contents in order to construct a query, and typically reveal no relationship structure among the returned documents. Thus we can easily find needles in haystacks, but understanding the haystack itself remains a challenge.

One approach for addressing this problem is to provide the user with a small, structured set of documents that reflect in some way the content space of the collection (or possibly a sub-collection consisting of documents related to a query). In this work we consider structure expressed in "threads", i.e., singly-connected chains of documents. For example, given a corpus of academic papers, we might want to identify the most significant lines of research, representing each by a citation chain of its most important contributing papers. Or, in response to a search for news articles from a particular time period, we might want to show the user the most significant stories from that period, and for each such story provide a timeline of its major events.

We formalize collection threading as the problem of finding diverse paths in a directed graph, where the nodes correspond to items in the collection (e.g., papers), and the edges indicate relationships (e.g., citations). A path in this graph describes a thread of related items, and by assigning weights to nodes and edges we can place an emphasis on high-quality paths. A diverse set of high-quality paths then forms a cover for the most important threads in the collection.

To model sets of paths in way that allows for repulsion (and hence diversity), we employ the structured determinantal point process (SDPP) framework [1], incorporating $k$-DPP extensions to control the size of the produced threads [2]. The SDPP framework provides a natural model over sets of structures where diversity is preferred, and offers polynomial-time algorithms for normalizing the model and sampling sets of structures.

However, even these polynomial-time algorithms can be too slow when dealing with many real-world datasets, since they scale quadratically in the number of features. (In our experiments, the exact algorithms would require over 200 terabytes of memory.) We address this problem using random feature projections, which reduce the dimensionality to a manageable level. Furthermore, we show that this reduction yields a close approximation to the original SDPP distribution, proving the following theorem based on a result of Magen and Zouzias [3].

**Theorem 1.** *Let $P^k$ be the exact $k$-SDPP distribution on sets of paths, and let $\tilde{P}^k(Y)$ be the $k$-SDPP distribution after projecting the similarity features to dimension $d = O(\max\{k/\epsilon, (\log(1/\delta) + \log N)/\epsilon^2\})$, where $N$ is the total number of possible path sets. Then with probability at least $1 - \delta$,*

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

Finally, we demonstrate our model using two real-world datasets. The first is the Cora research paper dataset, where we extract research threads from a set of about 30,000 computer science papers. Figure 1

- Retrieval and Reasoning in Distributed Case Bases
- Cooperative Information Gathering: A Distributed Problem Solving Approach
- MACRON: An Architecture for Multi-agent Cooperative Information Gathering
- Research Summary of Investigations Into Optimal Design-to-time Scheduling
- Control Heuristics for Scheduling in a Parallel Blackboard System
- Partial Global Planning: A Coordination Framework for Distributed Hypothesis Formation
- 3 Distributed Problem Solving and Planning
- Designing a Family of Coordination Algorithms
- Quantitative Modeling of Complex Environments
- Introducing the Tileworld: Experimentally Evaluating Agent Architectures
- Auto-blocking Matrix-Multiplication or Tracking BLAS3 Performance with Source Code
- A Model and Compilation Strategy for Out-of-Core Data Parallel Programs
- Tolerating Latency Through Software-Controlled Prefetching in Shared-Memory Multiprocessors
- Designing Memory Consistency Models For Shared-Memory Multiprocessors
- Sparcle: An Evolutionary Processor Design for Large-Scale Multiprocessors
- Integrating Message-Passing and Shared-Memory: Early Experience
- Integrated Shared-Memory and Message-Passing Communication in the Alewife Multiprocessor
- Compiling for Shared-Memory and Message-Passing
- Compiling for Distributed-Memory Systems
- Distributed Memory Compiler Design for Sparse Problems

Figure 1: Two example threads (left and right) from running a 5-SDPP on the Cora dataset.

| | 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 1: Quantitative results on news text, measured by similarity to human summaries. a = January-June; b = July-December. Starred (*) entries are significantly higher than others in the same column at 99% confidence.

shows some sample threads pulled from the collection by our method. The second dataset is a multi-year corpus of news text from the New York Times, where we produce timelines of the major events over six month periods. We compare our method against multiple baselines using human-produced news summaries as references. Table 1 contains measurements of similarity to the human summaries for our approach ($k$-SDPP) versus clustering (CLS) and non-max suppression (NMX) baselines.

Future work includes applying the diverse path model to other applications, and studying the empirical tradeoffs between speed, memory, and accuracy inherent in using random projections.

# References


[1] Alex Kulesza and Ben Taskar. Structured determinantal point processes. In *Proc. Neural Information Processing Systems*, 2010.

[2] A. Kulesza and B. Taskar. k-DPPs: fixed-size determinantal point processes. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[3] A. Magen and A. Zouzias. Near optimal dimensionality reductions that preserve volumes. *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 523–534, 2008.