Pacific Northwest
NATIONAL LABORATORY
*Proudly Operated by* **Battelle** *Since 1965*

# High-performance Algorithms and Software for Clustering Based on Constrained Low-rank Approximations

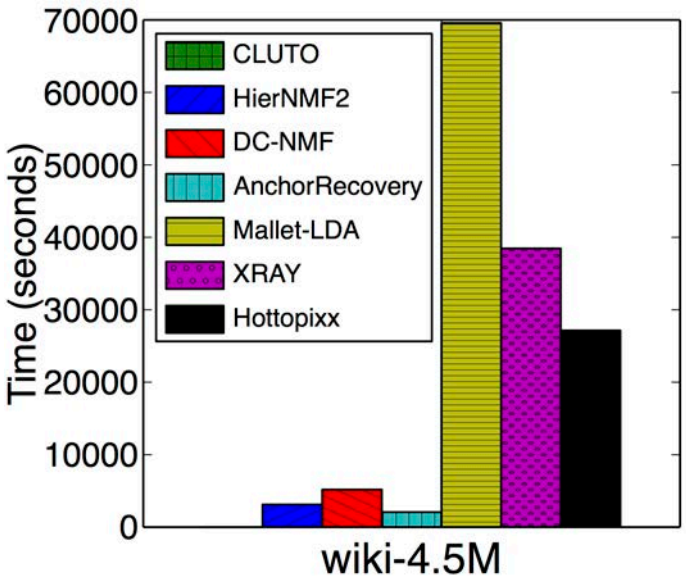
## CHALLENGE

Topic modeling from massive amounts of text data has been a growing research area for the last several years and is moving rapidly toward processing of streaming text data. The study of large-scale community detection within massive-scale network graphs is still in the early stages. A great opportunity exists to improve current community detection and topic modeling methods via the application of new, constrained low-rank approximation (CLRA) methods, which put both problems within a common algorithmic framework.

## CURRENT PRACTICE

The majority of applications for topic modeling currently use the latent Dirichlet allocation (LDA) algorithm. However, algorithms using CLRA demonstrate superior performance for topic modeling and clustering compared to some of the most commonly used methods available in open-source software packages, such as MALLET (LDA) and WEKA (K-means). While the CLRA methods take hours, LDA can take days, and K-means may not even finish. We illustrate these advantages in some recent benchmarks (featured in the figure at right), where software for 4.5 million Wikipedia documents was able to find 80 topics within 45 minutes on a 4-core commodity laptop using our methods. Meanwhile, LDA took about a day, Anchor Recovery produced poor-qualtity results, and CLUTO ran out of memory.

Designing scalable algorithms for large-scale problems and the ability to produce more accurate solutions faster in noisy real-life applications.




This figure demonstrates the performance of seven software packages with the number of topics = 80. LDA took nearly a day, CLUTO ran out of memory, AnchorRecovery was fast but produced poor-quality results (not shown). The CLRA methods are HierNMF2 and DC-NMF.

The purpose of community detection is to identify linkage-based or functional communities in a network. Earlier work in community detection mainly focused on finding linkage-based communities on small networks. Community detection on large networks is more challenging because many algorithms suitable for small networks are not scalable, and there is a dearth of ground-truth communities defined for large networks, rendering the analysis of results nearly impossible.

## TECHNICAL APPROACH

Our research aims to model text and graph analysis problems and design, verify, and deploy scalable numerical algorithms based on a powerful CLRA framework. The CLRA approach uses the nonnegative matrix factorization (NMF) method, which has demonstrated tremendous speed improvements for large-scale problems. For example, implementation of the popular multiplicative updating (MU) using a Hadoop system takes 50 minutes per iteration, whereas our algorithm takes less than one second on a cluster with 24 machines. We develop alternative NMF methods based on hierarchical factorizations for topic modeling, data clustering, graph clustering, and community detection, which are known to have inherent efficiencies, further improving the throughput on large-scale data. Many problems in data analytics can be formulated using NMF with various additional constraints and difference measures. The main advantages of these NMF formulations are flexibility for designing scalable algorithms and the ability to produce more accurate solutions in noisy real-life applications. We have designed a convergent algorithm, which is one of the fastest and is scalable to very large problems. However, as the model order (cluster dimension) becomes very large, even the fastest algorithm becomes slow and sometimes unaffordable. To resolve this issue, we design rank-2 NMF algorithms based on a divide-and-conquer scheme that is highly scalable to massive-scale problems.

## IMPACT

The CLRA framework addresses new issues and challenges, such as robustness, convergence to a global optimal solution, high scalability, and information fusion of content and structural information–all of which have the potential to uncover latent entity connections not discoverable from content or structure alone. If successful, these advances will push the frontiers of algorithm research into unexpected directions, benefiting a much wider class of application domains.

## Contacts

**Haesun Park**
Principal Investigator
(404) 933-6269
hpark@cc.gatech.edu

**David A. Bader**
(404) 385-4785
bader@cc.gatech.edu

**John R. Johnson**
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
*Proudly Operated by Battelle Since 1965*

Georgia Institute of Technology