SIAM Data Mining "Brings It" to Annual Meeting

Jeremy Kepner  
*Massachusetts Institute of Technology*

Sanjukta Bhowmick  
*University of Nebraska at Omaha*

Aydın Buluç  
*Lawrence Berkeley National Laboratory*

Rajmonda Caceres  
*Massachusetts Institute of Technology*

R. Jordan Crouser  
*Smith College*, jcrouser@smith.edu

*See next page for additional authors*

Follow this and additional works at: [https://scholarworks.smith.edu/csc_facpubs](https://scholarworks.smith.edu/csc_facpubs)

Part of the [Computer Sciences Commons](https://scholarworks.smith.edu/csc_facpubs)

**Recommended Citation**

Kepner, Jeremy; Bhowmick, Sanjukta; Buluç, Aydın; Caceres, Rajmonda; Crouser, R. Jordan; Gadepally, Vijay; Miller, Ben; and Webster, Jennifer, 'SIAM Data Mining "Brings It" to Annual Meeting' (2017). Computer Science: Faculty Publications, Smith College, Northampton, MA.  
[https://scholarworks.smith.edu/csc_facpubs/99](https://scholarworks.smith.edu/csc_facpubs/99)

This Article has been accepted for inclusion in Computer Science: Faculty Publications by an authorized administrator of Smith ScholarWorks. For more information, please contact scholarworks@smith.edu
Authors
Jeremy Kepner, Sanjukta Bhowmick, Aydın Buluç, Rajmonda Cáceres, R. Jordan Crouser, Vijay Gadepally, Ben Miller, and Jennifer Webster

This article is available at Smith ScholarWorks: https://scholarworks.smith.edu/csc_facpubs/99
SIAM Data Mining "Brings It" to Annual Meeting

Jeremy Kepner\(^1\) (Data Mining SIAG Chair), Sanjukta Bhowmick\(^2\), Aydın Buluç\(^3\), Rajmonda Caceres\(^1\), Jordan Crouser\(^4\), Vijay Gadepally\(^1\), Ben Miller\(^1\), Jennifer Webster\(^5\)

\(^1\)MIT Lincoln Laboratory, \(^2\)University of Nebraska Omaha, \(^3\)Lawrence Berkeley National Laboratory, \(^4\)Smith College, \(^5\)Pacific Northwest National Laboratory

The Data Mining Activity Group is one of SIAM’s most vibrant and dynamic activity groups. To better share our enthusiasm for data mining with the broader SIAM community, our activity group organized six minisymposia at the 2016 Annual Meeting. These minisymposia included 48 talks organized by 11 SIAM members on

- GraphBLAS (Aydın Buluç)
- Algorithms and statistical methods for noisy network analysis (Sanjukta Bhowmick & Ben Miller)
- Inferring networks from non-network data (Rajmonda Caceres, Ivan Brugere & Tanya Y. Berger-Wolf)
- Visual analytics (Jordan Crouser)
- Mining in graph data (Jennifer Webster, Mahantesh Halappanavar & Emilie Hogan)
- Scientific computing and big data (Vijay Gadepally)

These minisymposia were well received by the broader SIAM community, and below are some of the key highlights.

GraphBLAS
The theory of using matrices and vectors for graph computations has a long history, with a snapshot of the state-of-the-art being captured in the SIAM book *Graph Algorithms in the Language of Linear Algebra* by Kepner and Gilbert [1]. High-performance graph algorithms are often implemented with sparse matrices and linear algebra in many graph-processing systems. Example systems include the Combinatorial BLAS [2], D4M [3], GraphMat [4], and GPI [5]. The GraphBLAS.org [6] is a community initiative to standardize these different efforts to build a common foundation for graph algorithm developers. This minisymposium had 8 talks: Aydın Buluç from Lawrence Berkeley National Laboratory talked about the current status of the C language API and the

---

This material is based in part upon work supported by the NSF under grant number DMS-1312831, by DOE ASCR under contract number DE-AC02-05CH11231, and by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, recommendations and conclusions are those of the authors and are not necessarily endorsed by the United States Government.
ongoing efforts to develop a GraphBLAS-compliant parallel library in PGAS (partitioned global address space) languages. Jose Moreira and Manoj Kumar from IBM presented the Graph Programming Interface (GPI) as well as a proposal for a common binary format for storing graphs. Carl Yang from UC Davis talked about implementing breadth-first search utilizing the GraphBLAS primitives on clusters of GPU-equipped computers. Andrew Lumsdaine from Indiana University talked about the software and systems issues related to implementing the GraphBLAS Template Library (GBTL) [7] on different backends, such as CPUs and GPUs. Jeremy Kepner from MIT Lincoln Laboratory presented the mathematical foundations of the GraphBLAS [8], with an emphasis on incidence matrices as a preferred representation for graphs in databases. Scott McMillan from the CMU Software Engineering Institute dived deeper into the details of the GBTL library, with a focus on its frontend design. Narayanan Sundaram from Intel presented GraphMat (and its distributed cousin GraphPad), which is a highly-optimized graph library whose frontend is based on vertex programming and whose backend is based on linear algebra operations. Finally, Michael Wolf from Sandia National Laboratories presented miniTri [9], a triangle enumeration-based data analytics miniapp, with specific focus on a linear algebraic algorithm (though miniTri has alternative algorithms in it). The sessions had lively discussions during breaks and were attended by approximately 25 people.

Algorithms and Statistical Methods for Noisy Network Analysis

Dealing with errors and noise is a common problem that the network science research community is beginning to address. A two-part minisymposium demonstrated the diversity of approaches to this problem, focusing on statistical methods and algorithms for addressing issues arising from noise in networks. Several presentations discussed useful properties of networks, such as centrality metrics and connected components, and the ways noise in the observations can affect the analysis [10]. These talks included generative models for networks and statistically rigorous methods to estimate properties from sampled data [11]. Other talks focused on filtering techniques, such as using metadata to narrow a search from a cue vertex or emphasizing an interesting substructure [12]. Some speakers discussed how noise affects the analysis in specific disciplines including collaboration science [13], bioinformatics [14], and cybersecurity. The minisymposium concluded with a discussion among the speakers and the audience on the common themes that arose. Participants agreed that, as noisy network analysis continues to evolve as a subfield, addressing the lack of a common framework for modeling and quantifying noise is an exceptionally important challenge that would allow synthesis of related research in many diverse areas.

Inferring Networks from Non-Network Data

This minisymposium explored the important topic of network representation learning. In many practical settings, researchers are faced with having to make arbitrary decisions on how to construct networks from noisy, indirect, and diverse...
data. Papers presented on both sessions covered important highlights from the current state-of-art for this emerging research area. Several speakers discussed the importance of connecting the objective of a learning task, whether that is link prediction, diffusion estimation, or vertex classification, to the process of constructing and evaluating network representations [15-19]. Another important theme emphasized domain-specific notions of quality, for example, in the context of constructing robust correlation networks from biological and climate data [20, 21]. Overall, the minisymposium helped consolidate important ideas, insights, and perspectives aimed at developing a rigorous and cohesive framework for learning robust network representations.

**Scientific Computing and Big Data**

This two-part minisymposium was a great success. We had nine speakers from diverse organizations that shared their considerable experience working with scientific big data. In the first session, we heard from Dr. Vijay Gadepally (MIT), Dr. Siddharth Samsi (MIT), Dr. Manoj Kumar (IBM Research), Dr. Michel Kinsy (Boston University), and Dr. Shashank Yellapantula (GE Global Research). Dr. Gadepally and Dr. Samsi discussed advances in data management technologies [22–25], and Dr. Kumar presented a brief overview of a graph-based API IBM is developing [26]. Dr. Kinsy discussed a novel processing architecture for low power computations [27]. Finally, Dr. Yellapantula discussed GE’s big data problems and many potential areas of collaboration with the wider SIAM community [28]. During the second session, we heard from a number of people in the medical community. Dr. Ashok Krishnamurthy (RENCI, UNC Chapel Hill) presented their development of a large-scale clinical data warehouse at the University of North Carolina Health Center [29]. Dr. Steve Finkbeiner (UCSF, Gladstone Institute) presented his group’s development of new robotic sensors capable of generating terabytes of imaging data per day to better understand the affects and causes of amyotrophic lateral sclerosis (ALS) [30]. Dr. Andy Zimolzak (Harvard, Department of Veteran Affairs) discussed his group’s work in developing computational infrastructure for precision oncology [31]. Dr. Aaron Elmore (University of Chicago) concluded the second session by presenting a new tool his research team is developing to be the GitHub for data – DataHub [32]. The presentations were of great interest to the diverse audience and there were many interesting discussions during the two sessions. Overall, the speakers and participants were left with a greater understanding of some domain-specific problems and technical strategies for addressing such problems.

**Mining in Graph Data**

Our minisymposium on Mining in Graph Data started off with organizer Jennifer Webster presenting an overview of the topic. Dr. Webster covered some common issues, including the use of found data that can be messy and the bias introduced by translating real-world problems into the language of mathematics. She highlighted these issues with examples drawn from shipping networks. Following that presentation, a second organizer, Mahantesh Halappanavar, presented algorithms for large-scale community detection. In particular he
described his parallel implementation of the Louvain modularity maximization method. The convergence results showed close agreement with the serial implementation, but the speed-up on multiple processors was significant. His group tested graphs with up to 50 million vertices and 2 billion edges. This was joint work conducted with Ananth Kalyanaraman. Our third speaker, Jevin West, spoke on mining information from citation networks. He presented "the map equation," which is based on dynamics of movement in a network and is used to discover communities based on those dynamics. A demo of his software was presented, along with a discussion of how this method can be used to discover the time evolution of communities. The final speaker of the morning session was Kamesh Madduri, who discussed a matrix factorization method for evaluating network community structure. When given a graph and a set of communities, he uses a non-negative matrix factorization to discover the relative importance of communities. One advantage of this work is that it can accommodate overlapping communities. These four talks rounded out the morning session, and we had steady attendance around 40 in the audience for all talks.

The minisymposium continued in the afternoon with Dr. David Haglin discussing (in his words "ranting about") the many situations in which hyper-multi-graphs can be used and the current algorithmic and computational resource limitations to the analysis of such graphs. Dr. Haglin gave several examples of graphs in cyber and social networks, especially those where non-numeric edge information arises and where the graphs created become extremely large. Dr. Sanjukta Bhowmick then discussed her metrics for community permanence that aid in the mitigation of the noise present in real-world graphs. The permanence metric performed well across a variety of benchmark graphs and real-world data sets, and showed the stability of communities. We then saw Dr. Robert Bridges’ use of graph analysis techniques in the location of anomalous cyber activity as well as the more friendly changes in American football conferences. Dr. Bridges’ methods dealt with time-varying graphs, noisy data, and a host of other challenges in the generation, creation, and analysis of these graphs. The final talk of the minisymposium, given by Ariful Azad, was on comparing communities across graphs. When given two related graphs with communities identified, one might ask how the communities compare across those graphs. Azad gave examples of graphs created from biological data, such as MRI scans and also image segmentation over time. The Mixed Edge Cover (MEC) algorithm was used to match corresponding communities, and experimental results were given in these example data sets to show algorithm performance.

**Visual Analytics**

This minisymposium was organized by Prof. R. Jordan Crouser of Smith College. Visual analytics is “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas & Cook, 2006)[33] and is rapidly gaining ground as an important discipline complementary to applied mathematics. The two-session series featured speakers from Smith College, WPI, Bucknell, DePaul University, Washington University, MIT Lincoln Laboratory, and IBM Research, and covered
topics ranging from the design and evaluation of visual analytics systems to the role of human perception in data analysis.

References


[19] Brugere, I., Latent Graph Inference and Validation, *IEEE Inter. Conf. on Data Mining Workshop (ICDMW)*, 2015.


