HYPERION RESEARCH

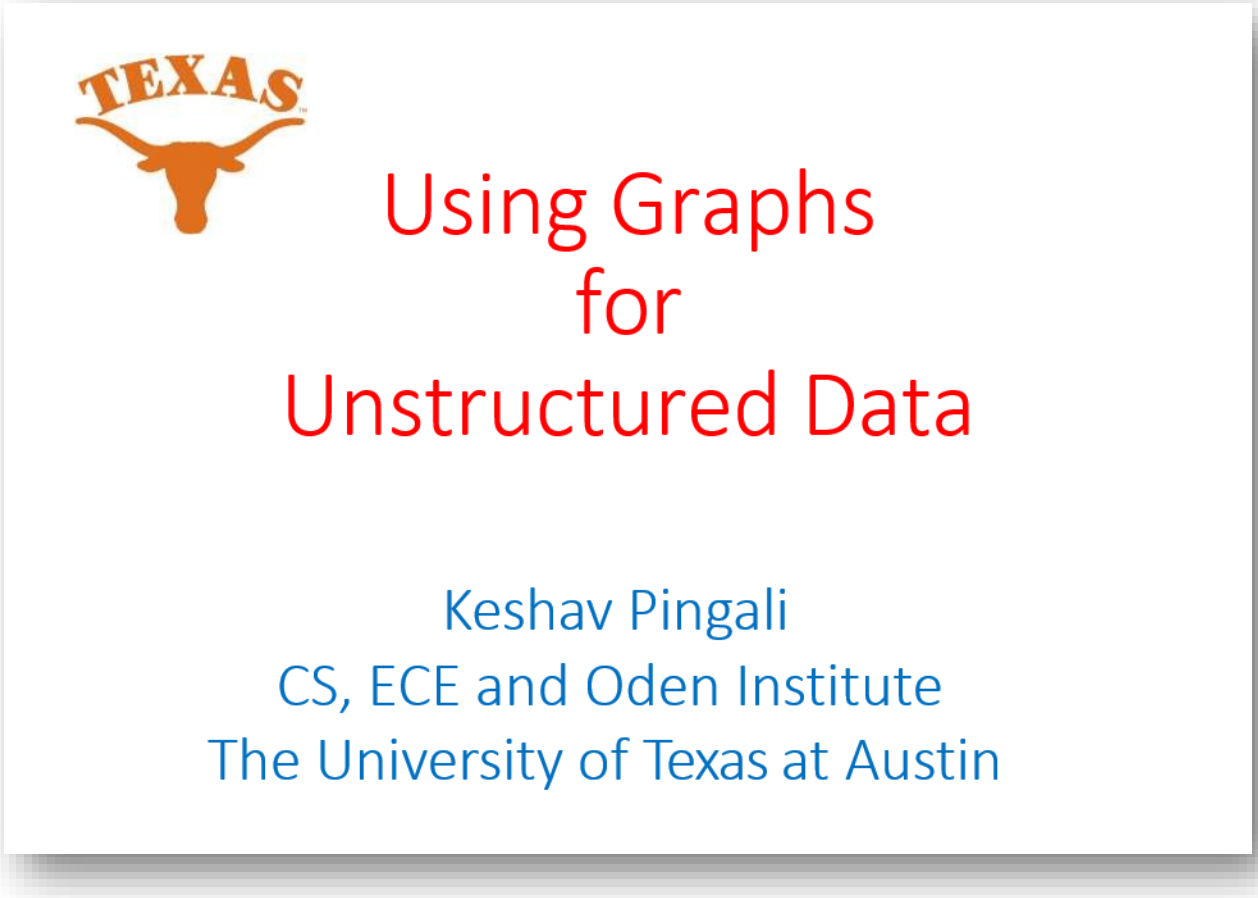
HPC User Forum Update

# Using Graphs for Unstructured Data

Thomas Gerard and Bob Sorensen
September 2020

## IN THIS UPDATE

The HPC User Forum was established in 1999 to promote the health of the global HPC industry and address issues of common concern to users. In September 2019, the 73rd HPC User Forum took place in Chicago, Illinois. This update summarizes a presentation from that meeting entitled, *Using Graphs for Unstructured Data*, given by Keshav Pingali, William Moncrief Chair of Grid and Distributed Computing, University of Texas at Austin.



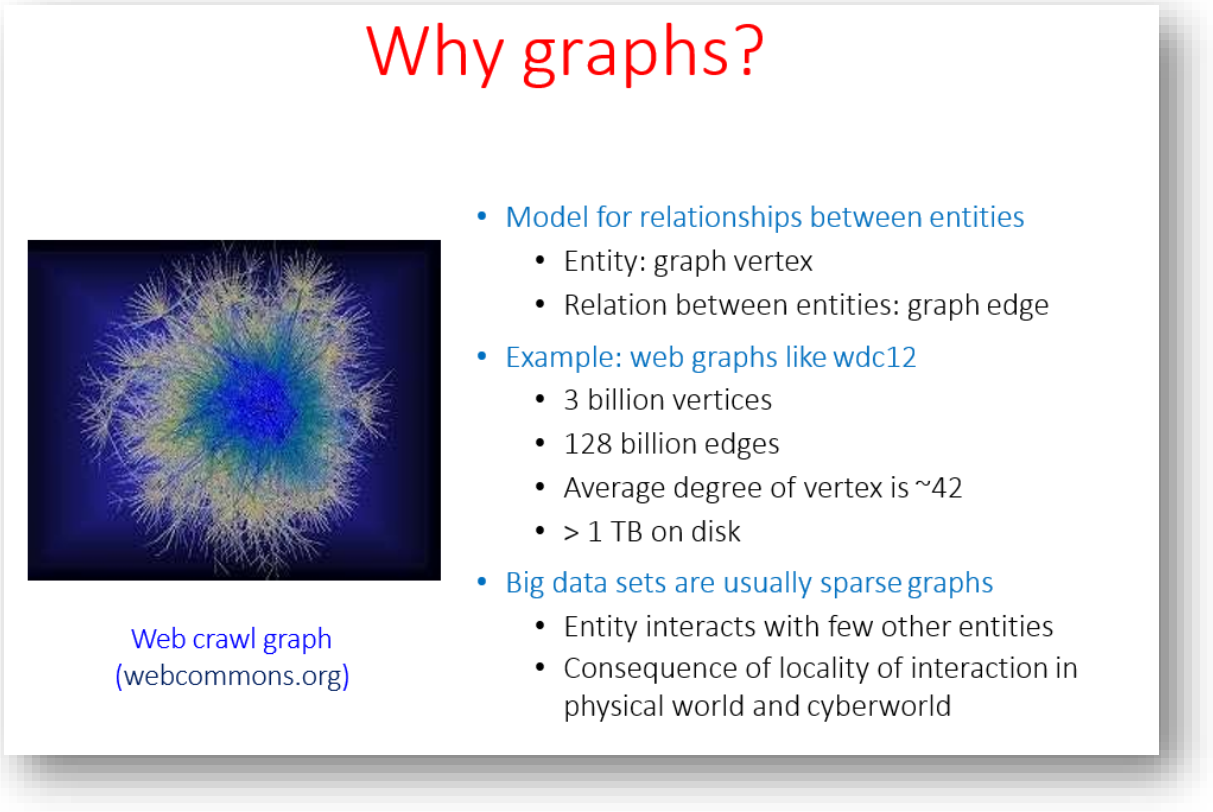

Source: UT Austin and Hyperion Research, 2020

## PRESENTATION: *USING GRAPHS FOR UNSTRUCTURED DATA*, KESHAV PINGALI, WILLIAM MONCRIEF CHAIR OF GRID AND DISTRIBUTED COMPUTING, UNIVERSITY OF TEXAS AT AUSTIN

Keshav Pingali, speaking on behalf of the CS, ECE, and Oden Institute at the University of Austin, discussed the ongoing projects of his Intelligent Software Systems Group, a group of 12 PhD students and postdocs funded by, among others, three major DARPA projects and several from NSF. One project is called the Galois system for parallel programming of unstructured problems. The group has also done work on adaptive control systems for principled accuracy/energy tradeoff in many computations. More recently, they've been exploring the use of machine learning in systems software.

First, Pingali poses a question – why use graphs for unstructured data? He posits that graphs are excellent models for understanding relationships between entities. The entities get mapped to graph vertices, then, if there is a relationship between two entities, that is represented by an edge – multiple edges representing multiple relationships between entities (called a multi-graph). In many of their problems, there are labels on nodes and labels on edges, all of which must be computed, leading to much interesting algorithm work in this area.

### FIGURE 1

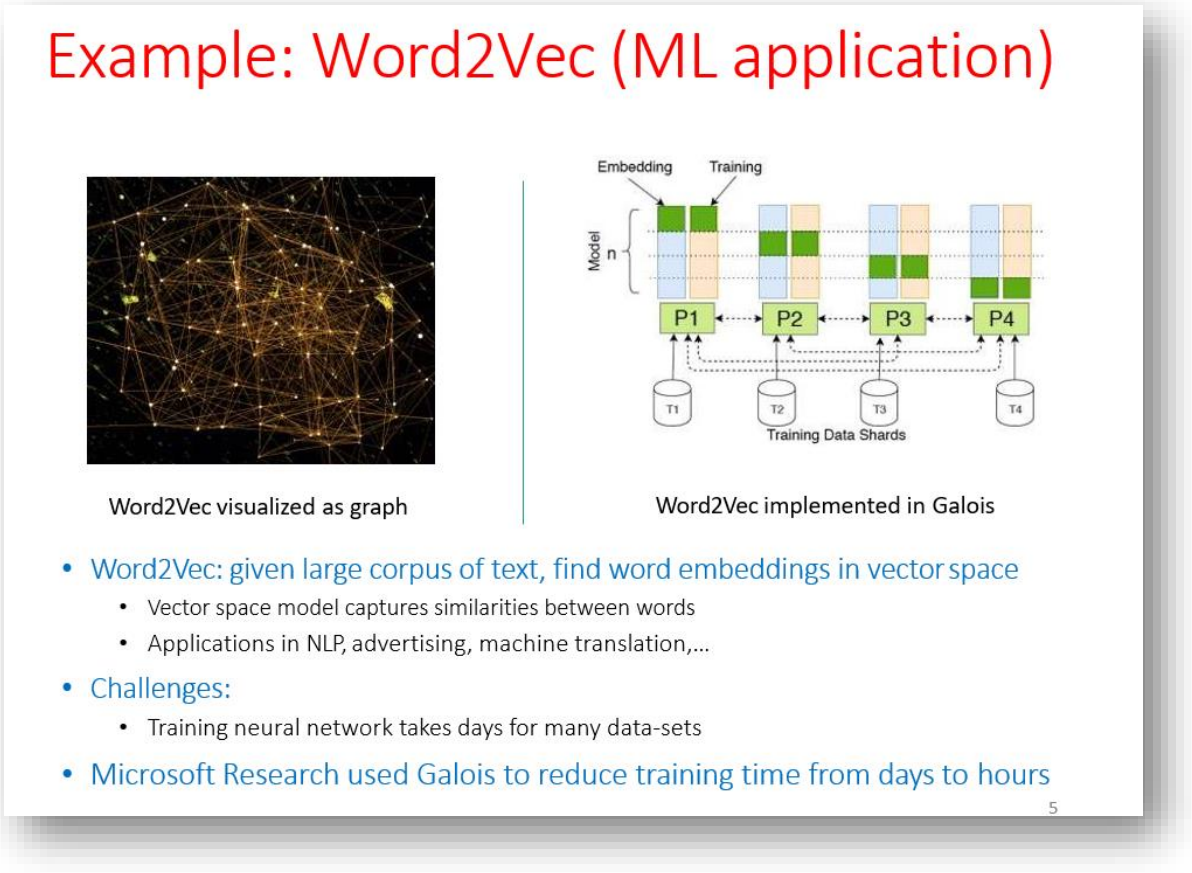



Source: UT Austin and Hyperion Research, 2020

Pingali presented a web crawl graph. He explained that with 3 billion vertices and 128 billion edges, the enormous graph offers little information to the naked eye. This type of graph, produced by companies like Google and Microsoft, is built by, 'scraping the web'. Each node represents a document on the web and if there is a hyperlink between two documents that is represented by an edge. This, he explained, is how search engines do page rank computations in order to discern which results to provide to a query. He points out that the average degree of a vertex is ~42 and occupies >1 TB of disk space. Pingali characterizes the graphs they work on as, "very large," and, "very sparse," as is common with big data sets.

To Illustrate the ubiquitous nature of graphs and their usefulness, Pingali discussed the variety of use cases for the Galois system. While they are effective in many fields, he chose to zero in on three: machine learning, engineering design and simulation, and security. In each example, vast sets of nodes, whether representing users or files, are linked together with edges which represent interactions, like reading or sending a message. This can be used to map and explore large sets of data, recognize patterns and "leaders" in social media environments, and even aid in the construction and refinement of meshes for traditional simulation, modelling, and graphics.

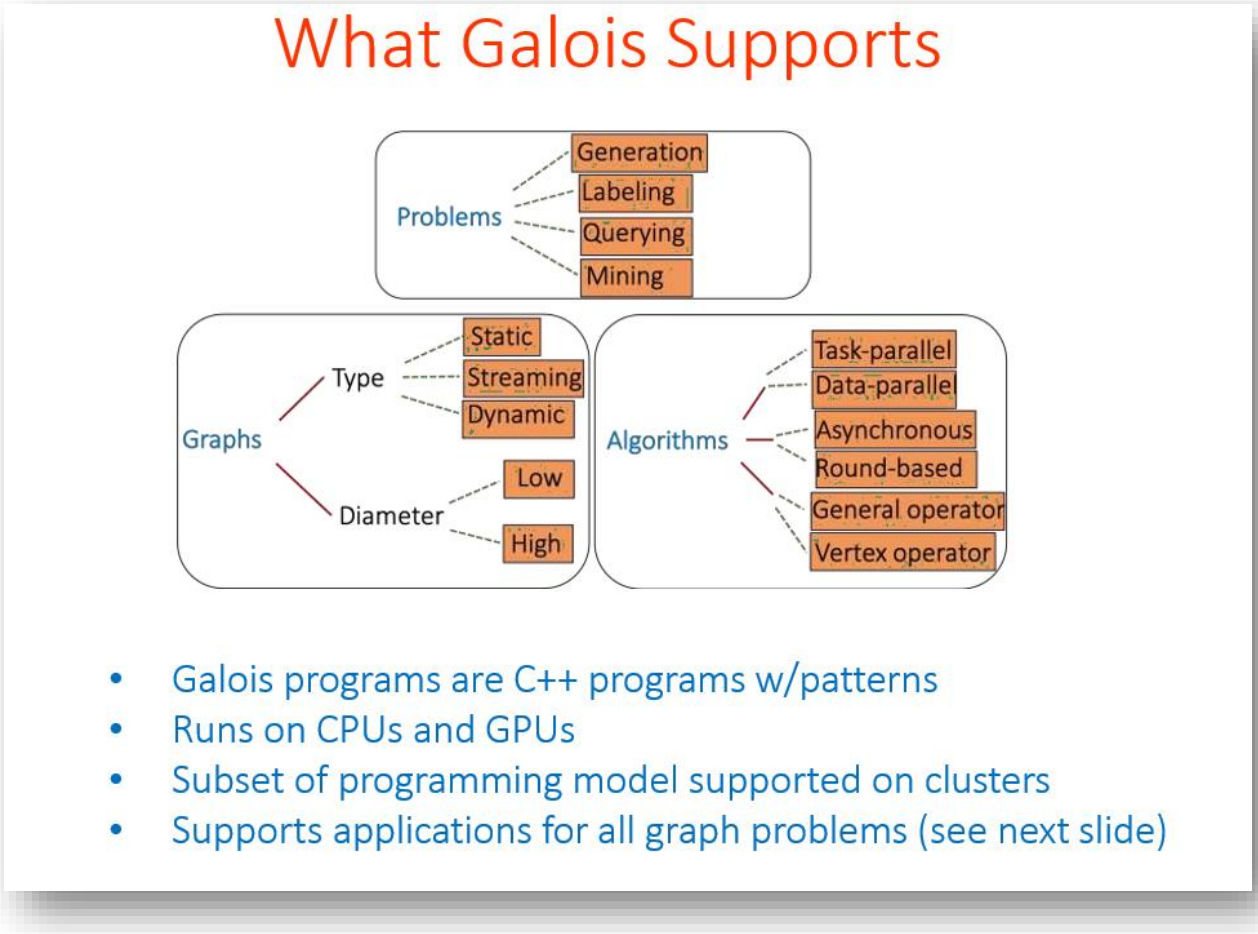
## FIGURE 2




Source: UT Austin and Hyperion Research, 2020

Moving on from his examples of real-world applications, Pingali went into, "abstractions," and how to categorize these graph computations. He divides graph problems into four categories: generation, labeling, querying, and mining. Graph generation is the construction of a graph by the addition and subtraction of vertices and edges. Labeling problems are concerned with the computation of the attributes of the nodes and edges, while the graph is invariant. A querying problem, also using an invariant graph structure, is used to recognize paths within graphs and is used in their intrusion detection application. The final problem, which Pingali notes for its compute intensity, is mining. This is for finding patterns and their frequencies in graphs. Once the problem is identified, other features must also be categorized: whether the graph is static, streaming, or dynamic, and its diameter, which is important when determining the algorithm that will most efficiently interact with the graph. Pingali points out that Galois programs are all C++ and supports an extremely wide variety of graph problems, types, and algorithms.

## FIGURE 3

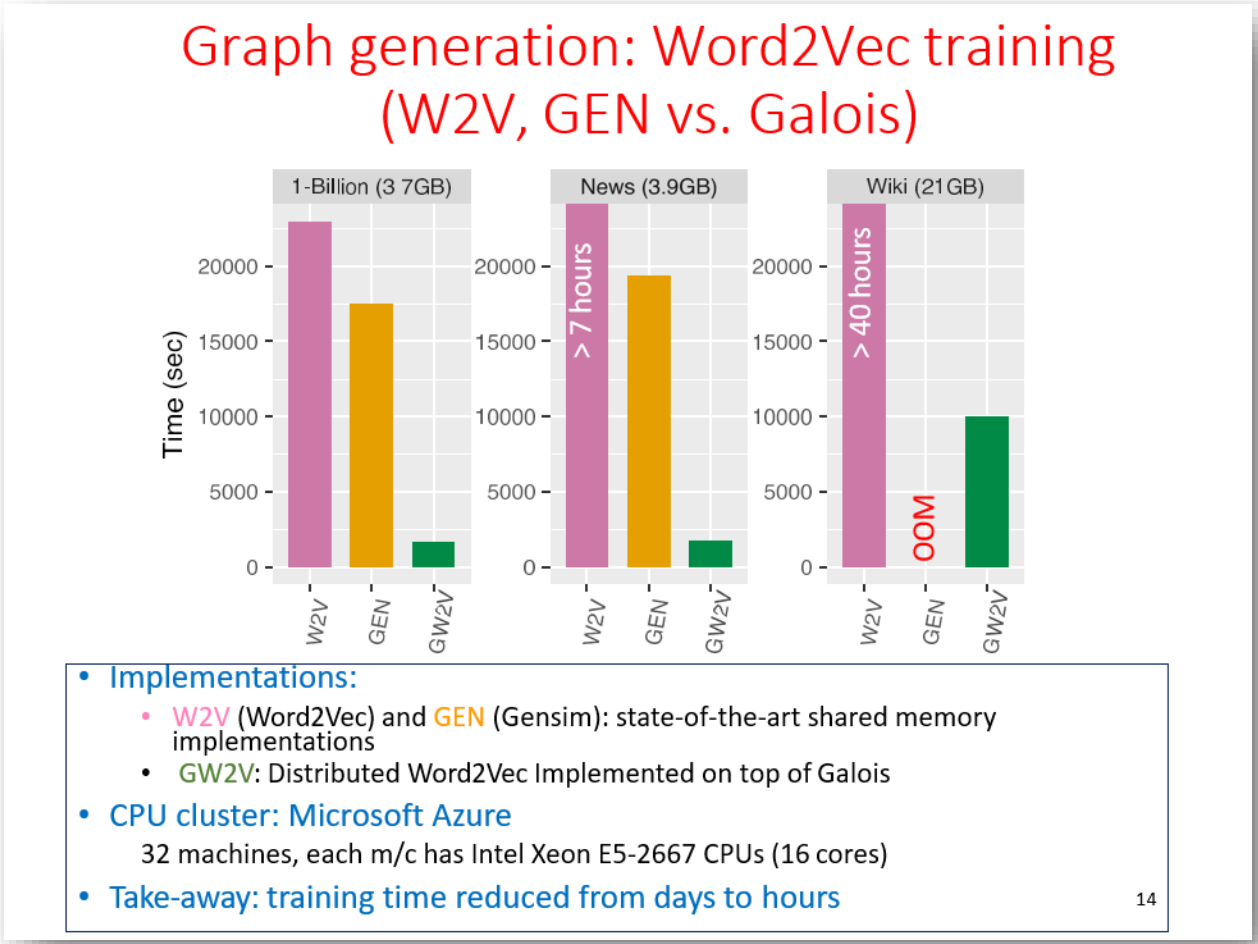



Source: UT Austin and Hyperion Research, 2020

Galois is a graph engine that understands graph data structures and how to compute with graphs. The Galois team have implemented numerous APIs like OpenCypher and GraphBLAS for mining, analytics,

querying, and the like. Pingali shows performance numbers for the aforementioned applications. In the work with Microsoft on Word2Vec in generating a similarities graph using all the content on Wikipedia, Galois reduced the training time from over 40 hours to under 3 hours. In demonstrating a labelling problem, he also points out that the Galois system works well on both CPUs (e.g., Skylake) and GPUs (e.g., NVIDIA), with GPU efficiency becoming more significant on more compute intensive applications. He also referenced using Intel's Optane as an alternative to using a big cluster just for memory.

## FIGURE 4




Source: UT Austin and Hyperion Research, 2020

Pingali closed by highlighting the ubiquity of graphs and the great diversity and complexity of their algorithms and by offering to discuss Galois and his work with anyone who was interested.

*For more information or to view this and other presentations given at HPC User Forums dating back to 2008, visit [www.hpcuserforum.com](http://www.hpcuserforum.com).*

## About Hyperion Research, LLC

Hyperion Research provides data driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

## Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com

---