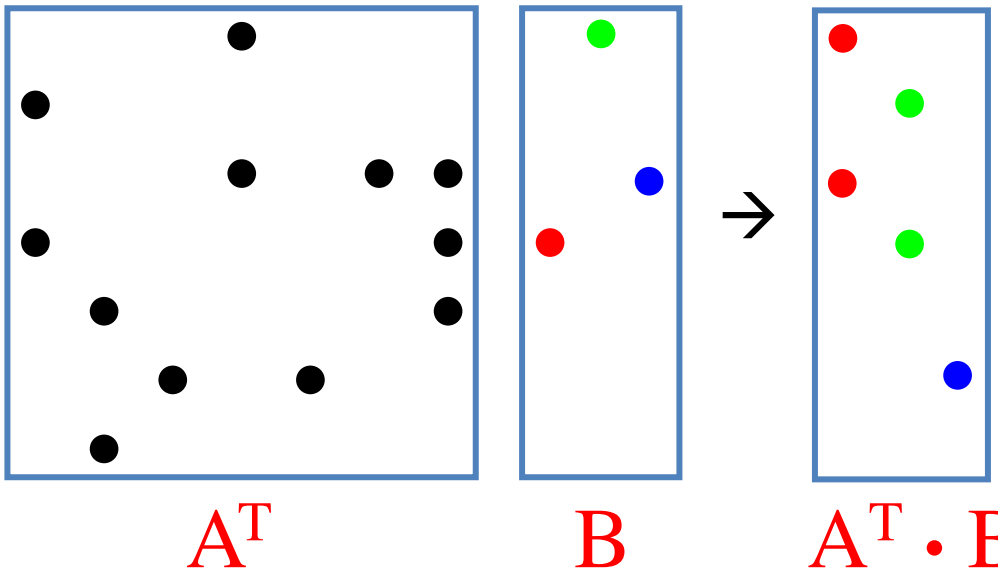
Berkeley
UNIVERSITY OF CALIFORNIA

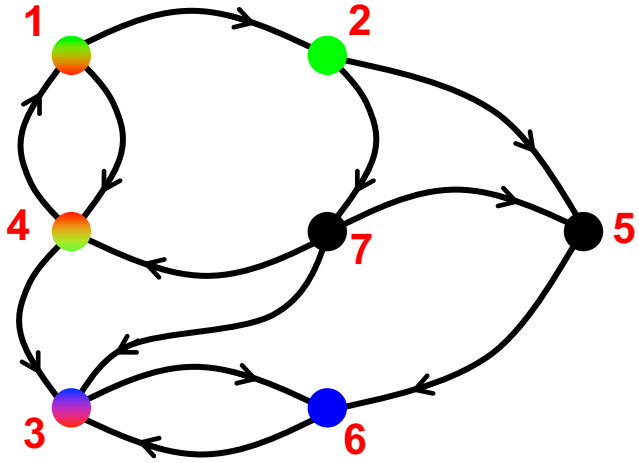
GraphBLAST: A high-performance C++ GPU library implementing GraphBLAS

Aydın Buluç
Computational Research Division, LBNL
EECS Department, UC Berkeley

ECP Annual Meeting, Feb 5, 2020

Graphs in the language of matrices





- Sparse array representation => space efficient
- Sparse matrix-matrix multiplication => work efficient
- Three possible levels of parallelism: searches, vertices, edges
- Highly-parallel implementation for Betweenness Centrality*

*: A measure of influence in graphs, based on shortest paths

The GraphBLAS effort

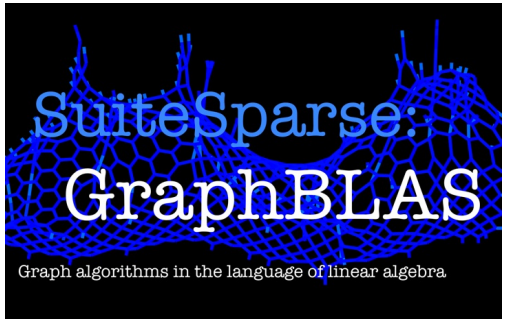
Standards for Graph Algorithm Primitives

Tim Mattson (Intel Corporation), David Bader (Georgia Institute of Technology), Jon Berry (Sandia National Laboratory), Aydin Buluc (Lawrence Berkeley National Laboratory), Jack Dongarra (University of Tennessee), Christos Faloutsos (Carnegie Melon University), John Feo (Pacific Northwest National Laboratory), John Gilbert (University of California at Santa Barbara), Joseph Gonzalez (University of California at Berkeley), Bruce Hendrickson (Sandia National Laboratory), Jeremy Kepner (Massachusetts Institute of Technology), Charles Leiserson (Massachusetts Institute of Technology), Andrew Lumsdaine (Indiana University), David Padua (University of Illinois at Urbana-Champaign), Stephen Poole (Oak Ridge National Laboratory), Steve Reinhardt (Cray Corporation), Mike Stonebraker (Massachusetts Institute of Technology), Steve Wallach (Convey Corporation), Andrew Yoo (Lawrence Livermore National Laboratory)

Abstract-- It is our view that the state of the art in constructing a large collection of graph algorithms in terms of linear algebraic operations is mature enough to support the emergence of a standard set of primitive building blocks. This paper is a position paper defining the problem and announcing our intention to launch an open effort to define this standard.

- The GraphBLAS Forum: http://graphblas.org
- Graphs: Architectures, Programming, and Learning (GrAPL @IPDPS): http://hpc.pnl.gov/grapl/

SuiteSparse::GraphBLAS



- From Tim Davis (Texas A&M)
- First conforming implementation of C API
- Features [1]:
  - 960 semirings built in; also user-defined semirings
  - Fast incremental updates using non-blocking mode and “zombies”
  - Several sparse data structures & polyalgorithms under the hood
- Already multithreaded [2]
- Performance on graph benchmarks (e.g. triangles, k-truss) comparable to highly-tuned custom C code
- Included in Debian and Ubuntu Linux distributions
- Used as computational engine in commercial RedisGraph product

[1] Davis, Timothy A. "Algorithm 1000: SuiteSparse: GraphBLAS: Graph Algorithms in the Language of Sparse Linear Algebra." ACM Transactions on Mathematical Software (TOMS) 45.4 (2019): 44.
[2] Aznaveh, Mohsen, et al. "Parallel GraphBLAS with OpenMP." CSC20, SIAM Workshop on Combinatorial Scientific Computing. SIAM. 2020.

GraphBLAS C API Spec (http://graphblas.org)

- **Goal:** A crucial piece of the GraphBLAS effort is to translate the mathematical specification to an actual Application Programming Interface (API) that
  i. is faithful to the mathematics as much as possible, and
  ii. enables efficient implementations on modern hardware.
- **Impact:** All graph and machine learning algorithms that can be expressed in the language of linear algebra
- **Innovation:** Function signatures (e.g. mxm, vxm, assign, extract), parallelism constructs (blocking v. non-blocking), fundamental objects (masks, matrices, vectors, descriptors), a hierarchy of algebras (functions, monoids, and semiring)

```
GrB_info GrB_mxm(GrB_Matrix          *C,      // destination
                 const GrB_Matrix     Mask,
                 const GrB_BinaryOp   accum,
                 const GrB_Semiring   op,
                 const GrB_Matrix     A,
                 const GrB_Matrix     B
                [, const Descriptor   desc]);
```

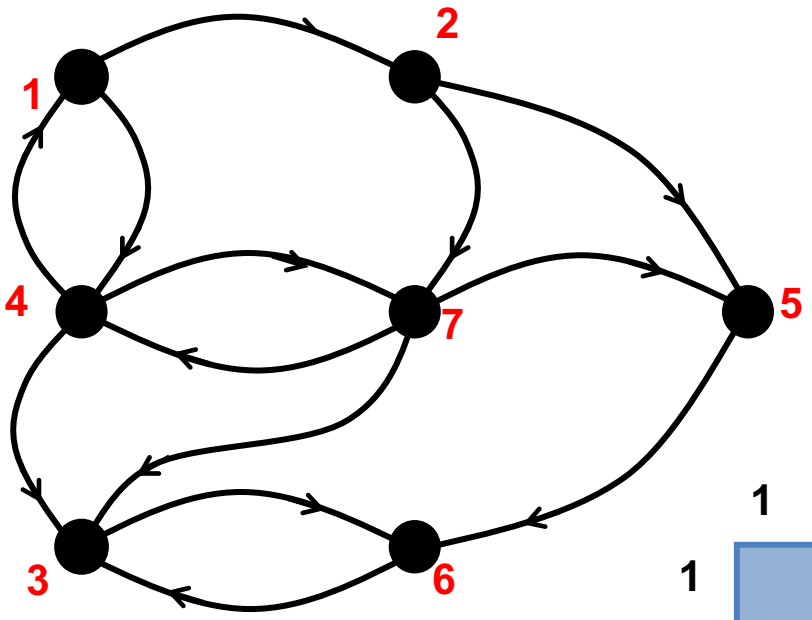
$$C(\neg M) \oplus= A^T \oplus.\otimes B^T$$

A.Buluç, T. Mattson, S. McMillan, J. Moreira, C. Yang. "The GraphBLAS C API Specification", version 1.3.0

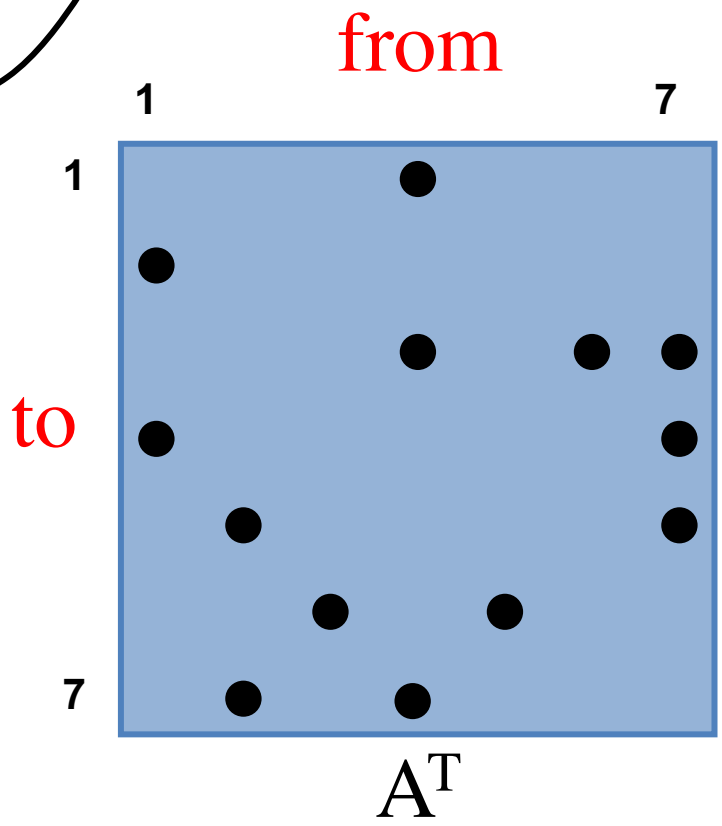
Examples of semirings in graph algorithms

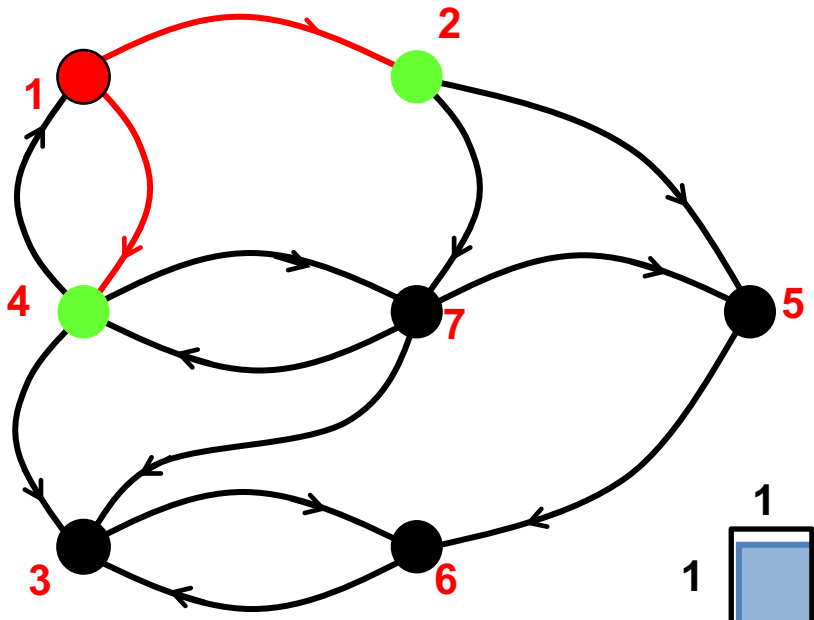
Real field: **(R, +, x)**	Classical numerical linear algebra
Boolean algebra: **({0 1}, |, &)**	Graph connectivity
Tropical semiring: **(R U {∞}, min, +)**	Shortest paths
(S, select, select)	Select subgraph, or contract nodes to form quotient graph
(edge/vertex attributes, vertex data aggregation, edge data processing)	Schema for user-specified computation at vertices and edges
(R, max, +)	Graph matching &network alignment
(R, min, times)	Maximal independent set

- **Shortened semiring notation: (Set, Add, Multiply)**. Both identities omitted.
- **Add:** Traverses edges, **Multiply:** Combines edges/paths at a vertex
- Neither add nor multiply needs to have an inverse.
- Both **add** and **multiply** are **associative**, **multiply distributes** over **add**



Breadth-first search in the language of matrices

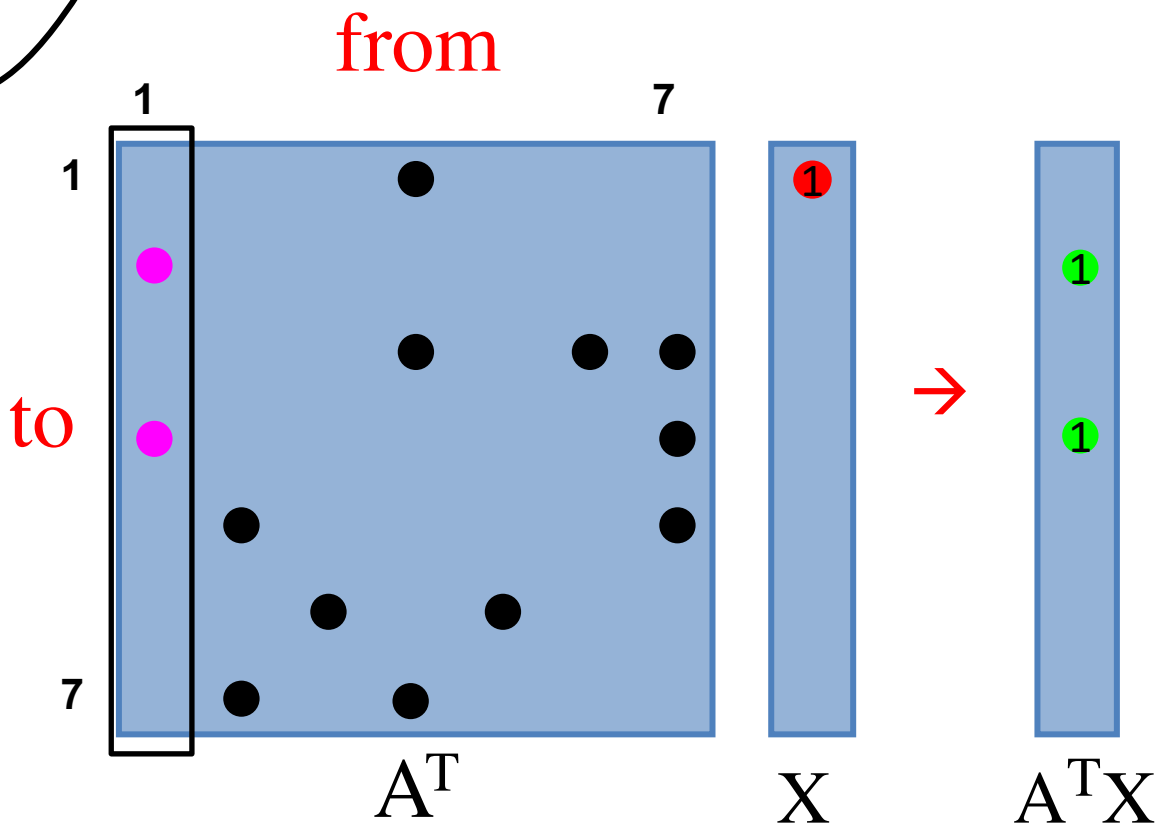




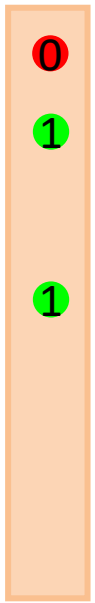
Particular semiring operations:

Multiply: select2nd

Add: minimum



parents:



Input sparsity

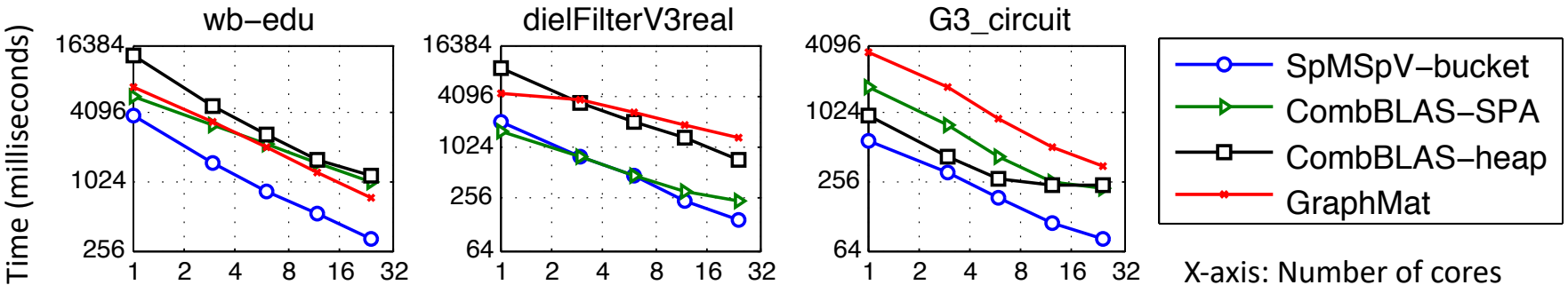
- What was the cost of that $A^T x$ in the previous slide?
- If x is dense, it is O(nnz(A)) = O(m) where m=#edges
- **If x is sparse**, it is

$$\sum_{i:x_i \neq 0} nnz(A_{i:})$$

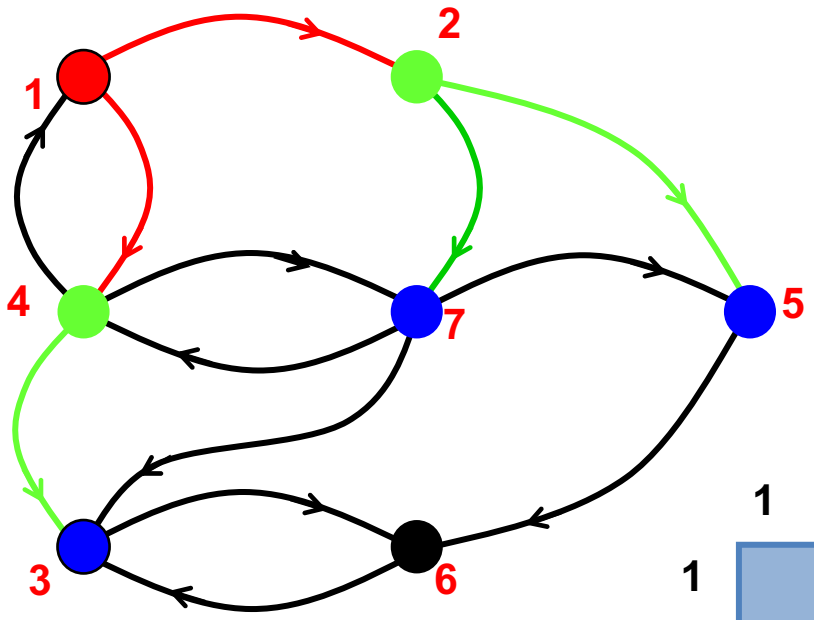
- ***Over all iterations*** of BFS, the cost sums up to O(nnz(A)), because no x_i appears twice in the input.
- Note that this is ***optimal*** for conventional (top-down) BFS
- Many people outside the community miss this* observation and ***mistakenly think*** SpMV based BFS is suboptimal by a factor of the graph diameter.

A work-efficient parallel algorithm for sparse matrix-sparse vector multiplication (SpMSpV)

- **Moral:** You should use this algorithm for exploiting input (vector) sparsity in SpMV in multicore and many-core architectures
- **Algorithmic innovation:**
  - Attains work-efficiency by arranging necessary columns of the matrix into buckets where each bucket is processed by a single thread
  - Avoids synchronization by row-wise partitioning of the matrix on the fly
- **Performance:**
  - First ever work-efficient algorithm for SpMSpV that attains up to 15x speedup on a 24-core Intel Ivy Bridge processor and up to 49x speedup on a 64-core KNL processor
  - Up to an order of magnitude faster than its competitors, especially for sparser vector

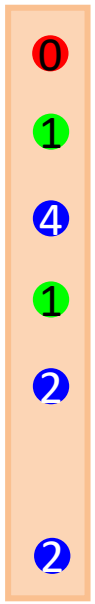


A.Azad, A. Buluç. A work-efficient parallel sparse matrix-sparse vector multiplication algorithm. IPDPS'17



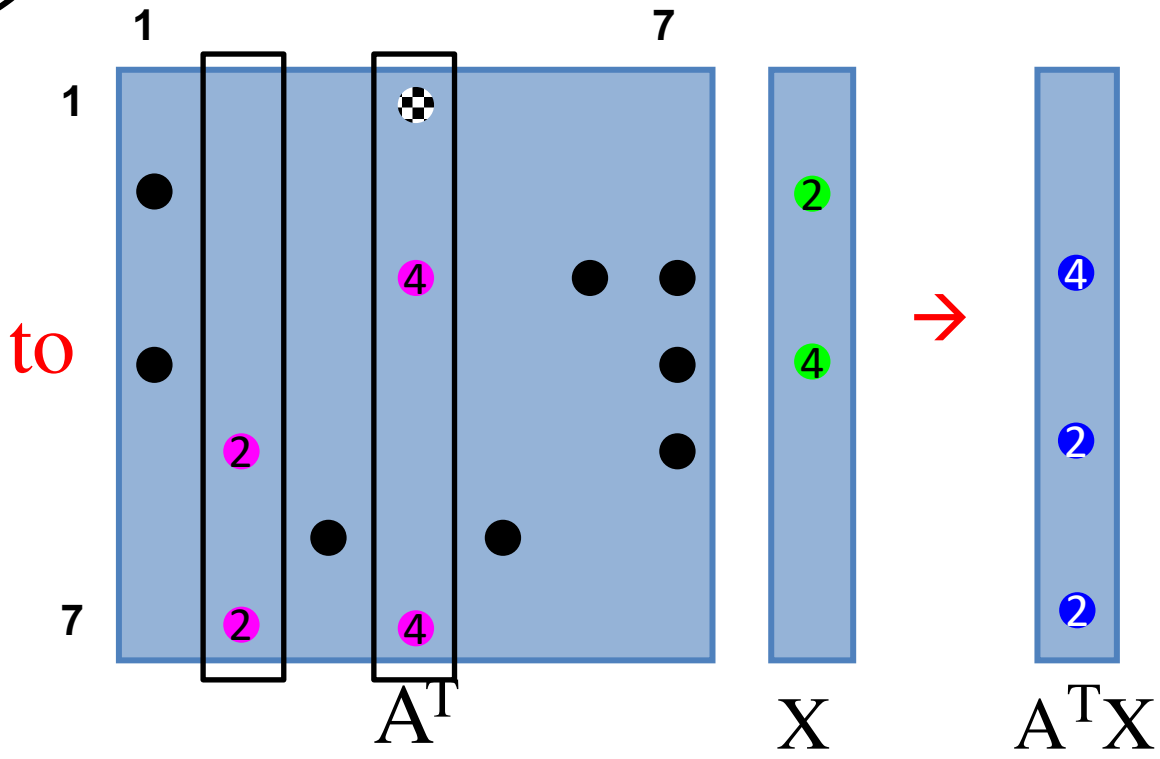
Select vertex with
minimum label as parent

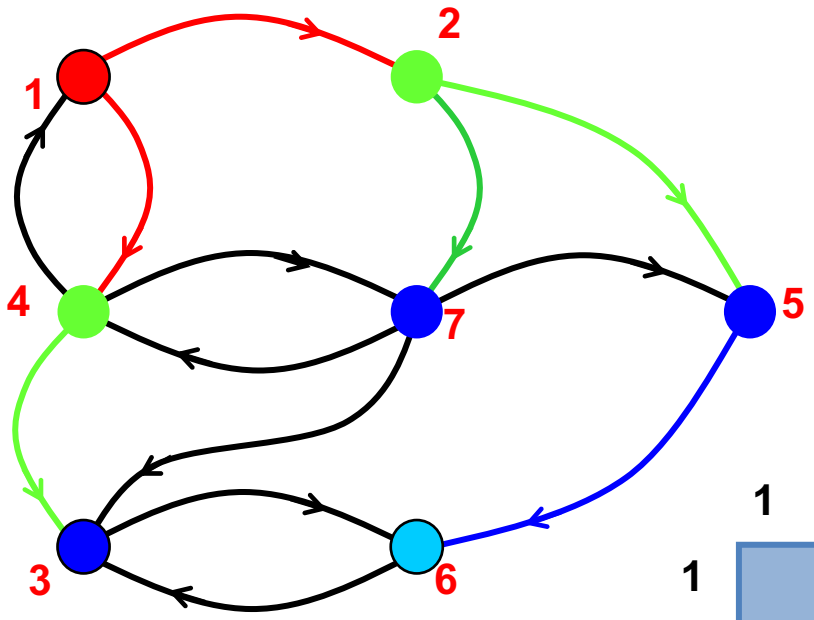
parents:



from

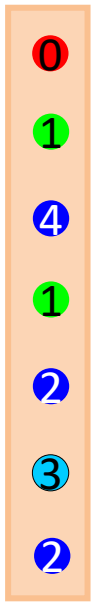
to

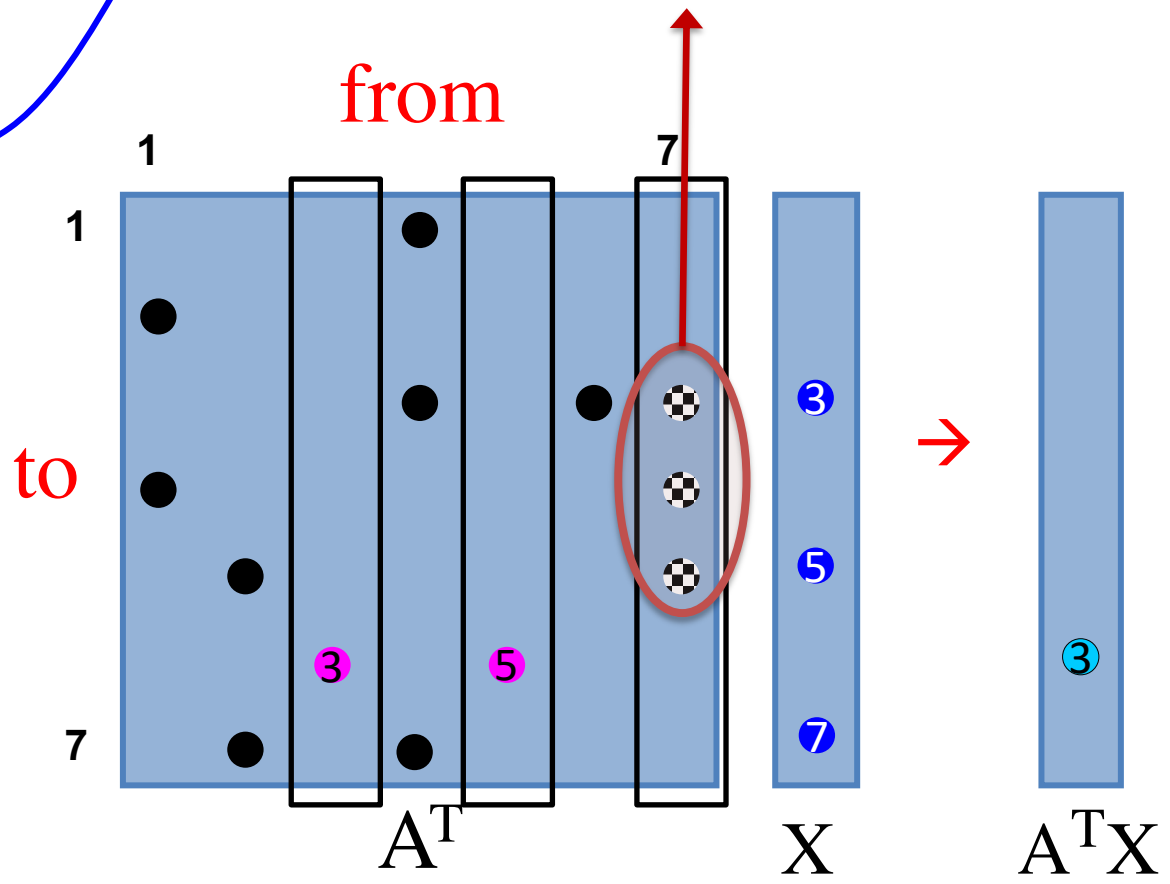


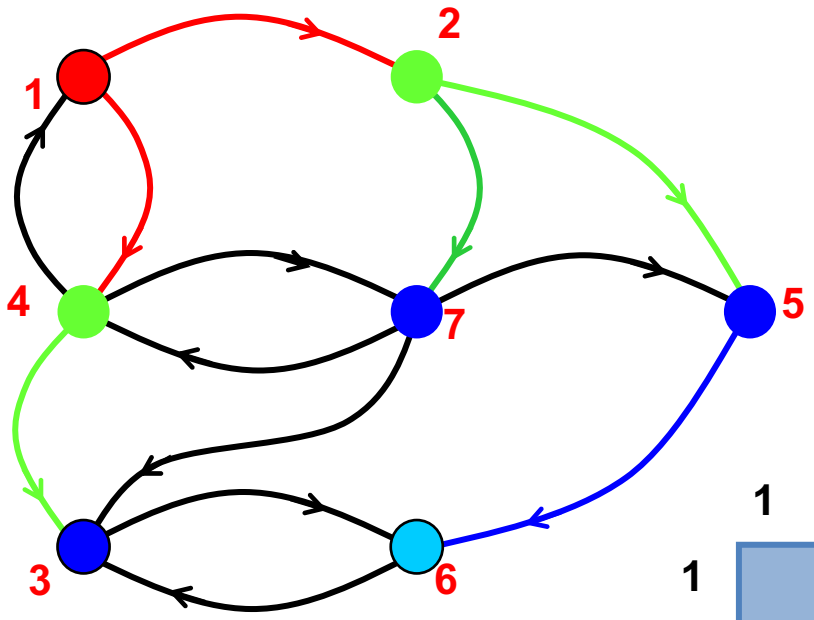


- Masks avoid formation of temporaries and can enable automatic direction optimization
- These footballs are nonzeros that are ***masked out*** by the parents array

parents:

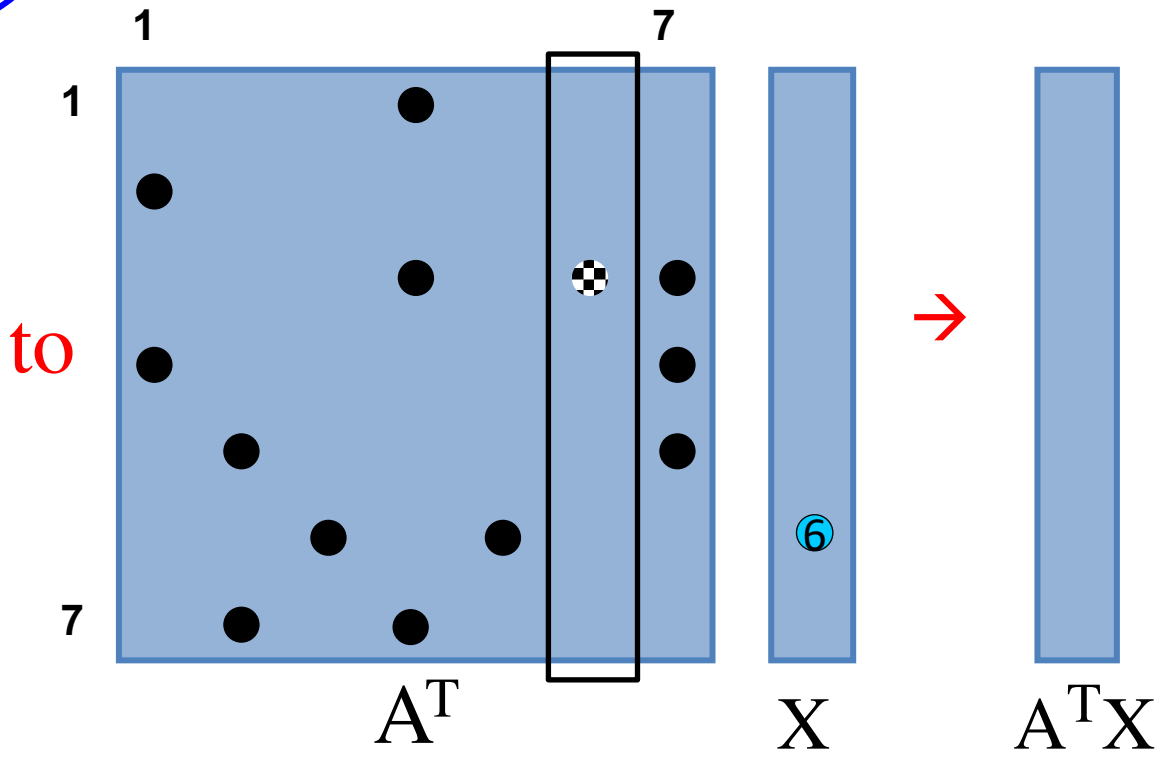






from

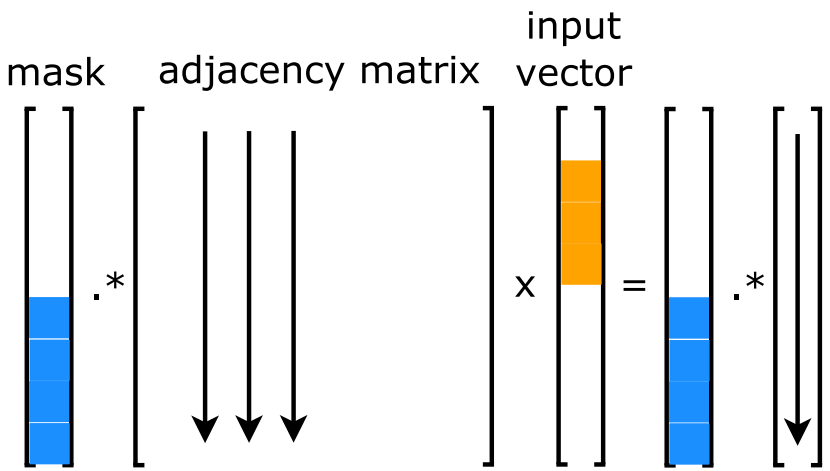
to



Output sparsity via masks

- The actual operation is $x = A^T x$.* p

  p is the parents array and .* is elementwise multiplication
- At first, our vision was limited: we only thought about eliminating temporaries in GrB_mxv
- But it was important enough to motivate the inclusion of masks into the GraphBLAS spec, though in limited form



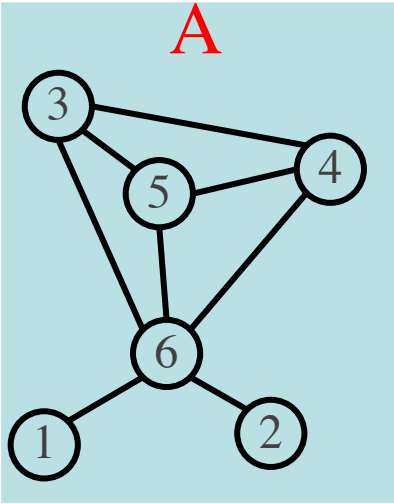
Column-based matvec w/ mask

Idea was to run the same column-based algorithm, but checking against a mask before writing to output

BFS in GraphBLAS with Masks

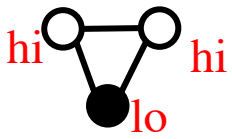
```
GrB_Vector q;                                   // vertices visited in each level
GrB_Vector_new(&q,GrB_BOOL,n);                  // Vector<bool> q(n) = false
GrB_Vector_setElement(q,(bool)true,s);          // q[s] = true, false everywhere else

GrB_Monoid Lor;                                 // Logical-or monoid
GrB_Monoid_new(&Lor,GrB_LOR,false);

GrB_Semiring Boolean;                           // Boolean semiring
GrB_Semiring_new(&Boolean,Lor,GrB_LAND);

GrB_Descriptor desc;                            // Descriptor for vxm
GrB_Descriptor_new(&desc);
GrB_Descriptor_set(desc,GrB_MASK,GrB_SCMP);     // invert the mask
GrB_Descriptor_set(desc,GrB_OUTP,GrB_REPLACE);// clear the output before assignment

GrB_UnaryOp apply_level;
GrB_UnaryOp_new(&apply_level,return_level,GrB_INT32,GrB_BOOL);

/*
 * BFS traversal and label the vertices.
 */
level = 0;
GrB_Index nvals;
do {
  ++level;                                      // next level (start with 1)
  GrB_apply(*v,GrB_NULL,GrB_PLUS_INT32,apply_level,q,GrB_NULL);  // v[q] = level
  GrB_vxm(q,*v,GrB_NULL,Boolean,q,A,desc);      // q[!v] = q ||.&& A ; finds all the
                                                // unvisited successors from current q
  GrB_Vector_nvals(&nvals, q);
} while (nvals);                                // if there is no successor in q, we are done.
```

Counting triangles

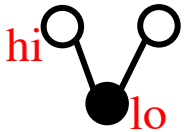


Thanks to triangle counting, we knew or sensed that something smarter algorithmically could be done than just eliminating temporaries.

Cohen's algorithm to count triangles:



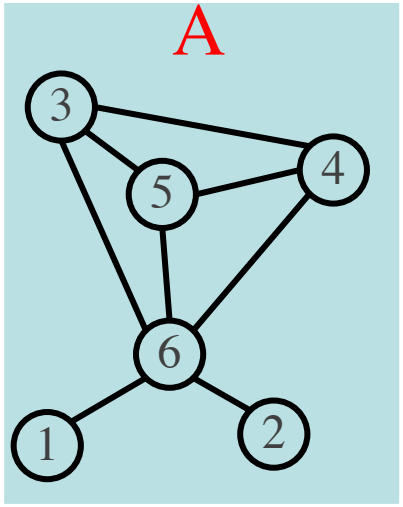
- Count triangles by lowest-degree vertex.



- Enumerate "low-hinged" wedges.

- Keep wedges that close.

Counting triangles

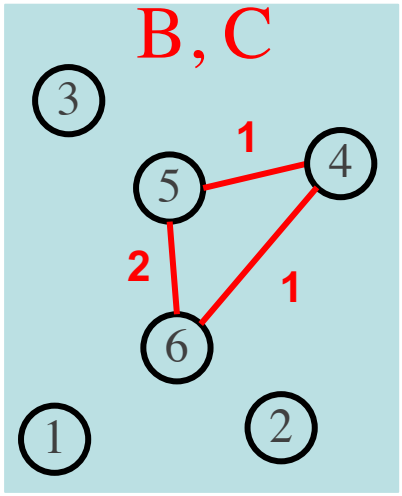


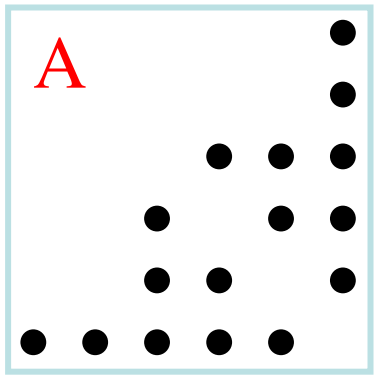
$A = L + U$ (hi->lo + lo->hi)

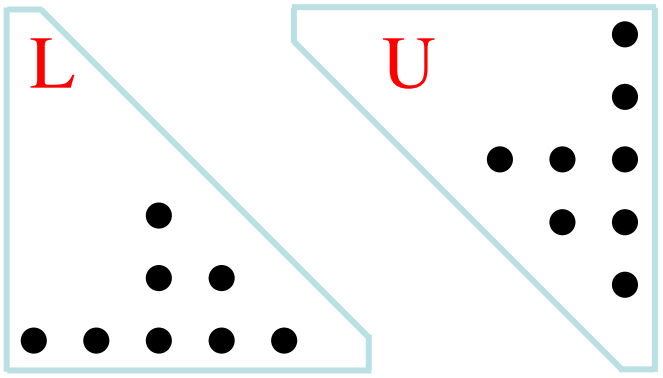
$L \times U = B$ (wedge, low hinge)

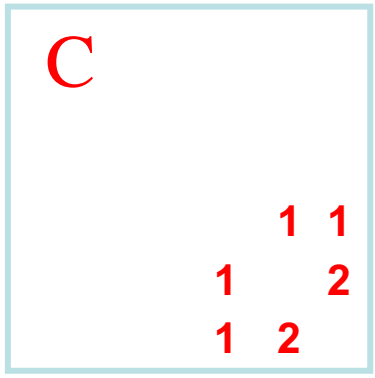
$A \wedge B = C$ (closed wedge)

$sum(C)/2 =$ **4 triangles**





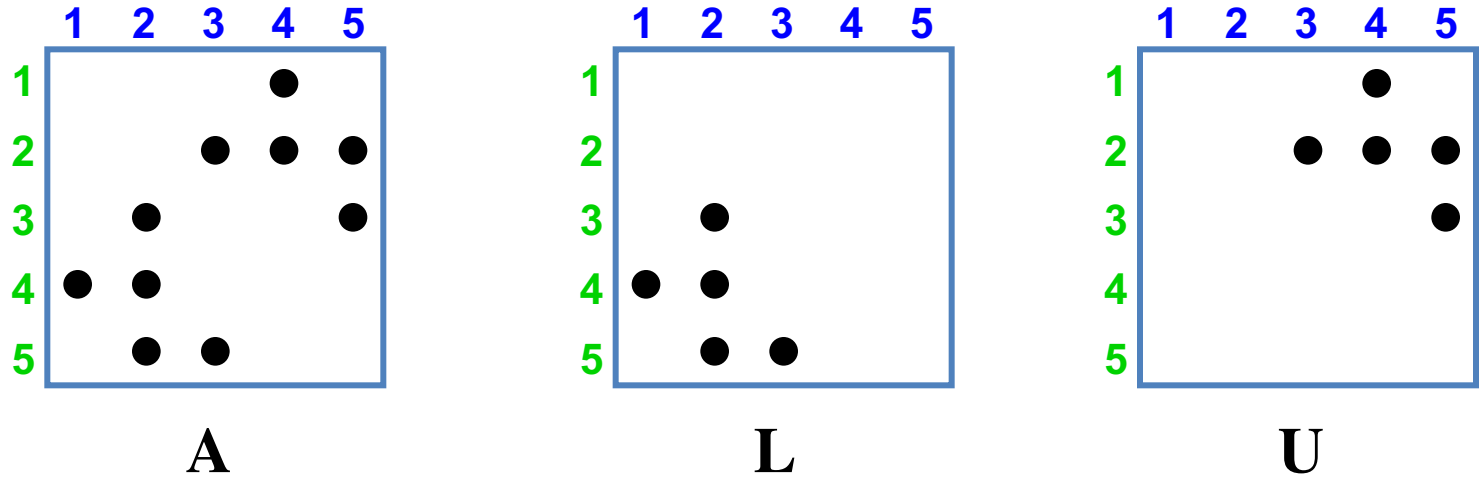




Azad, B., Gilbert. "Parallel triangle counting and enumeration using matrix algebra". *IPDPSW, 2015*

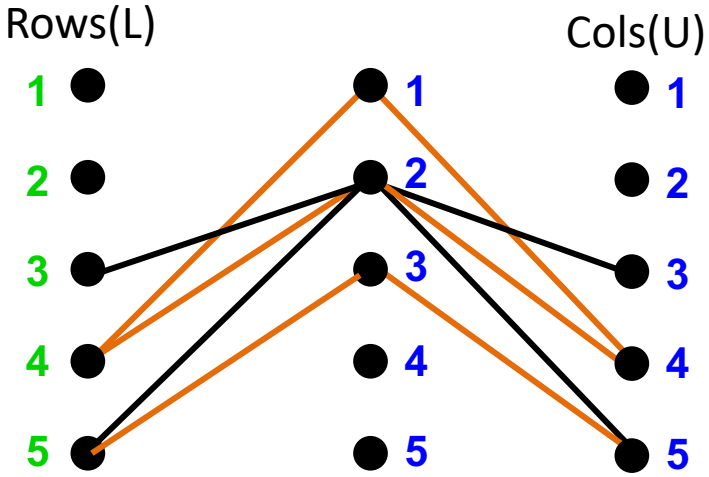
The first paper that has the word "Masked SpGEMM" in it

Exploiting Masks to avoid computation



Triangle counting example: **B = (LU) ^ A** ⇔ **mxm (&B, A, GrB_NULL, Int32AddMul, L, U)**

- Orange edges can not contribute to the output, so drop them before computation
- If the mask is really sparse, just run the inner product SpGEMM on mask nonzeros
- The inner product algorithm is *equivalent to the set intersection algorithm* Andrew L. wanted (at HPEC'17) GraphBLAS to beat



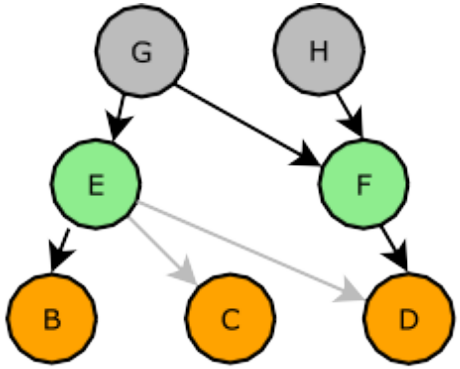
Triangle Counting in GraphBLAS

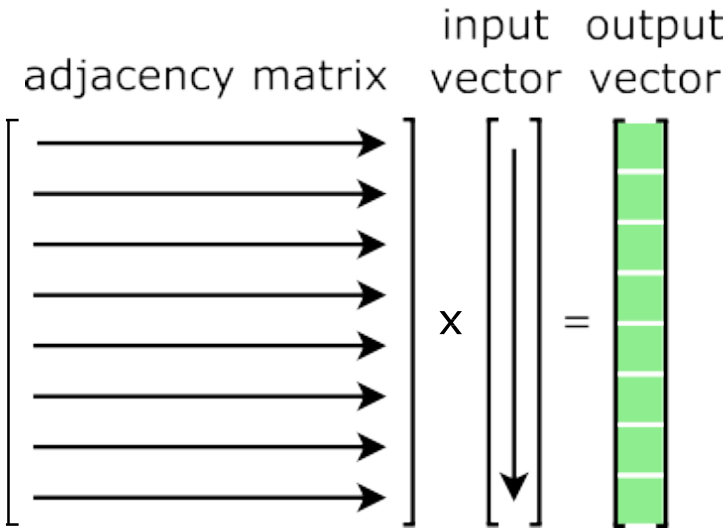
```
/*
 * Given, L, the lower triangular portion of n x n adjacency matrix A (of and
 * undirected graph), computes the number of triangles in the graph.
 */
uint64_t triangle_count(GrB_Matrix L)                  // L: NxN, lower-triangular, bool
{
  GrB_Index n;
  GrB_Matrix_nrows(&n, L);                             // n = # of vertices

  GrB_Matrix C;
  GrB_Matrix_new(&C, GrB_UINT64, n, n);

  GrB_Monoid UInt64Plus;                               // integer plus monoid
  GrB_Monoid_new(&UInt64Plus,GrB_PLUS_UINT64,0ul);

  GrB_Semiring UInt64Arithmetic;                       // integer arithmetic semiring
  GrB_Semiring_new(&UInt64Arithmetic,UInt64Plus,GrB_TIMES_UINT64);

  GrB_Descriptor desc_tb;                              // Descriptor for mxm
  GrB_Descriptor_new(&desc_tb);
  GrB_Descriptor_set(desc_tb,GrB_INP1,GrB_TRAN); // transpose the second matrix

  GrB_mxm(C, L, GrB_NULL, UInt64Arithmetic, L, L, desc_tb); // C<L> = L *.+ L'

  uint64_t count;
  GrB_reduce(&count, GrB_NULL, UInt64Plus, C, GrB_NULL);     // 1-norm of C

  GrB_free(&C);                          // C matrix no longer needed
  GrB_free(&UInt64Arithmetic);           // Semiring no longer needed
  GrB_free(&UInt64Plus);                 // Monoid no longer needed
  GrB_free(&desc_tb);                    // descriptor no longer needed

  return count;
}
```

http://graphblas.org

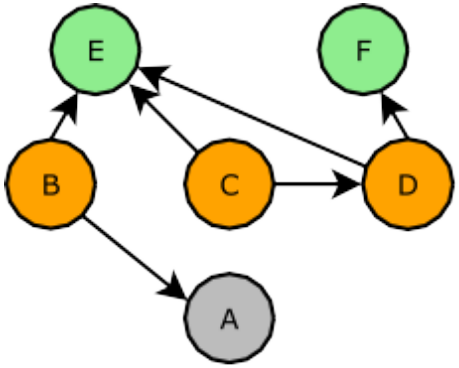
Push-pull ≡ column-row matvec

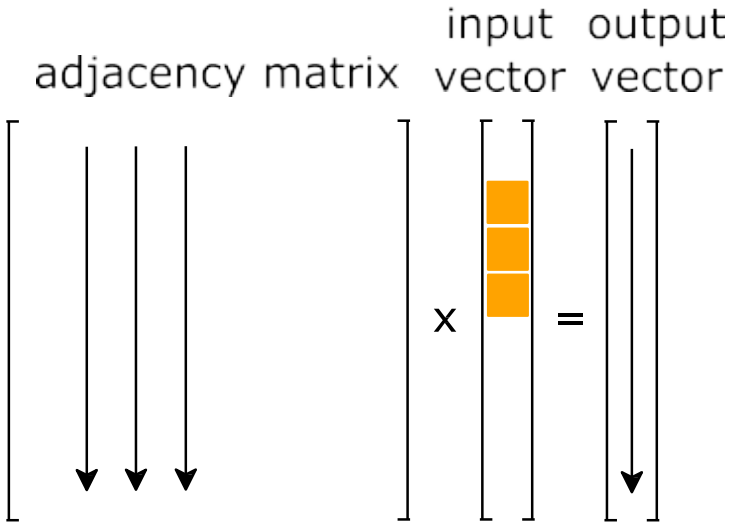
Pull





Push





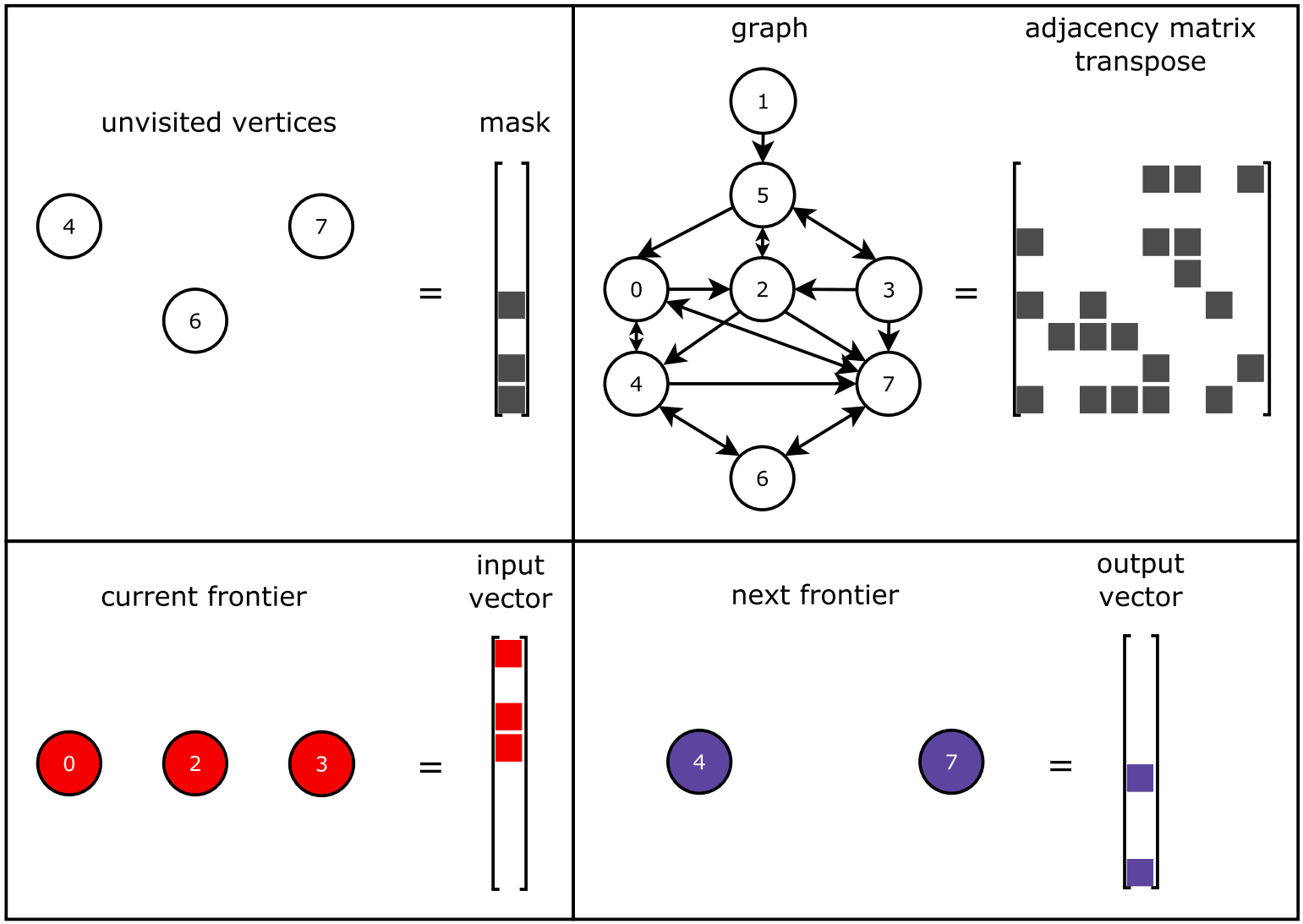
Yang, C., Buluc, A. and Owens, J.D., Implementing Push-Pull Efficiently in GraphBLAS. *ICPP'18*

Push-pull ≡ column-row matvec

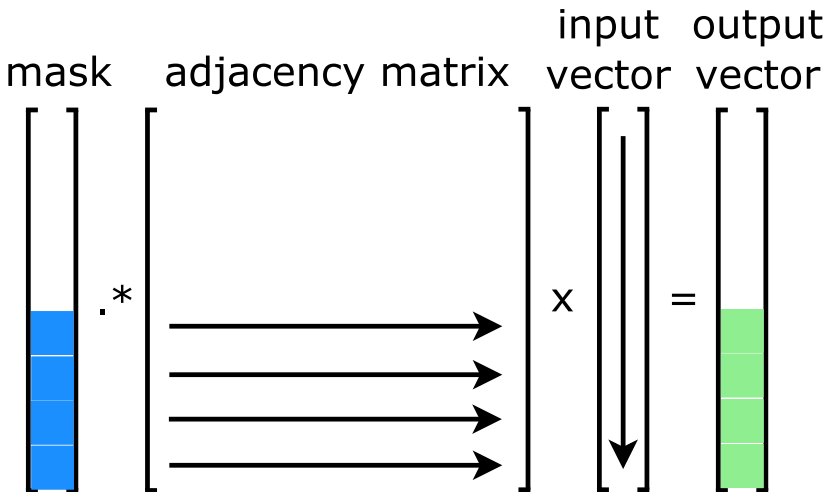
This is a story on how languages (and in this case APIs) change our thinking and drive our creative process

- Carl Yang and I pondered quite a bit on whether it was possible to implement direction optimization in the language of matrices *
- Push-pull (also known as direction optimization) was just about running a row- vs. column-based matvec
- But it wouldn't be competitive it its pure form because you were pulling from every vertex, not just unexplored ones.
- A year or so later, GraphBLAS had "masks"
- Now it was totally obvious how to make push-pull competitive in GraphBLAS

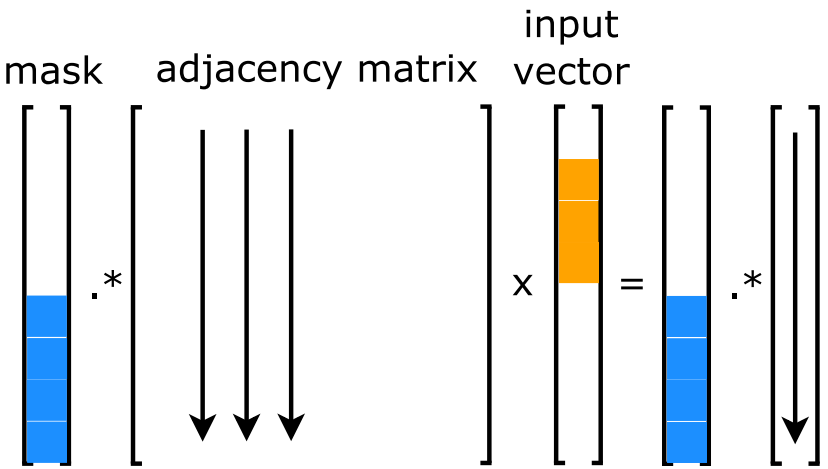
Enter “masks”


unvisited vertices
mask
4
7
6
=
graph
adjacency matrix
transpose
1
5
0
2
3
4
7
6
=
current frontier
input
vector
0
2
3
=
next frontier
output
vector
4
7
=

Masks make “pull” implementable competitively in GraphBLAS



Row-based matvec w/ mask



Column-based matvec w/ mask

- **Pull** is better for sufficiently sparse masks; **push** otherwise
- **Claim**: “direction optimization” would have been discovered automatically by the GraphBLAS runtime if we designed the interface back half a decade ago.

GraphBLAST

- First “high-performance” GraphBLAS implementation on the GPU
- Optimized to take advantage of both input and output sparsity
- Automatic direction-optimization through the use of masks
- Competitive with fastest GPU (Gunrock) and CPU (Ligra) codes
- Outperforms multithreaded SuiteSparse::GraphBLAS

Design principles:

1. Exploit input sparsity => direction-optimization
2. Exploit output sparsity => masking
3. Proper load-balancing => key for GPU implementations

Extensively evaluated on (more implemented, google for github repo)

- Breadth-first-search (BFS)
- Single-source shortest-path (SSSP)
- PageRank (PR)
- Triangle counting (TC)

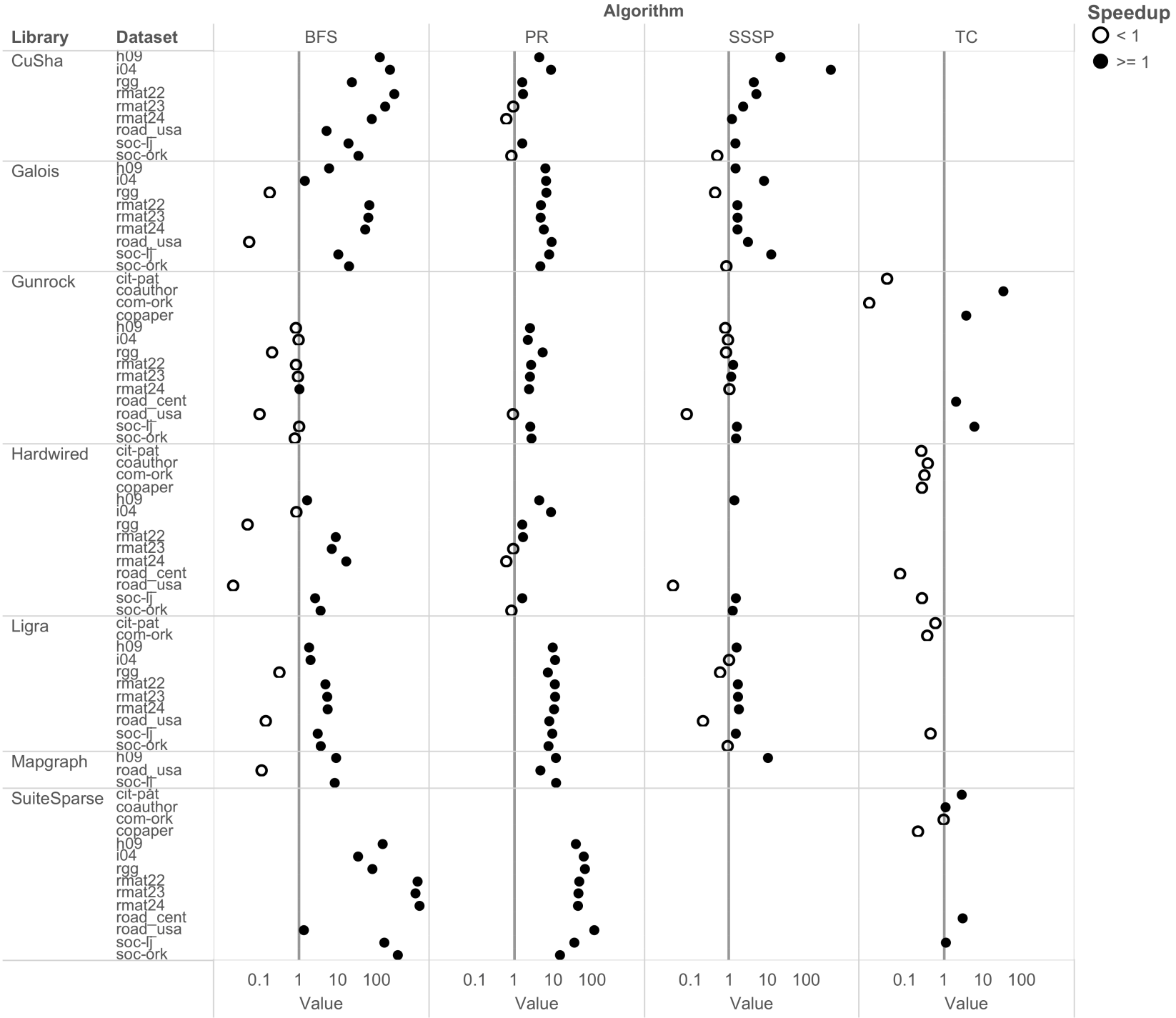
https://github.com/gunrock/graphblast

Yang, B., Owens, “GraphBLAST: A High-Performance Linear Algebra-based Graph Framework on the GPU”, arXiv

GraphBLAST syntax

```cpp
#include <graphblas/graphblas.hpp>

void bfs(Vector<float>*       v,
         const Matrix<float>* A,
         Index                s,
         Descriptor*          desc) {
  Index A_nrows;
  A->nrows(&A_nrows);
  float d = 1.f;

  Vector<float> f1(A_nrows);
  Vector<float> f2(A_nrows);
  std::vector<Index> indices(1, s);
  std::vector<float> values(1, 1.f);
  f1.build(&indices, &values, 1, GrB_NULL);

  v->fill(0.f);
  float c = 1.f;
  while (c > 0) {
    // Assign level d at indices f1 to visited vector v
    graphblas::assign(v, &f1, GrB_NULL, d, GrB_ALL, A_nrows, desc);
    // Set mask to use structural complement (negation)
    desc->toggle(GrB_MASK);
    // Multiply frontier f1 by transpose of matrix A using visited vector v as mask
    // Semiring: Boolean semiring (see Table 4)
    graphblas::vxm(&f2, v, GrB_NULL, LogicalOrAndSemiring<float>(), &f1, A, desc);
    // Set mask to not use structural complement (negation)
    desc->toggle(GrB_MASK);
    f2.swap(&f1);
    // Check how many vertices of frontier f1 are active, stop when number reaches 0
    // Monoid: Standard addition (see Table 4)
    graphblas::reduce(&c, GrB_NULL, PlusMonoid<float>(), &f1, desc);
    d++;
  }
}
```

- To avoid have many different descriptors that are different minimally, GraphBLAST introduces the convenience function desc::toggle.
- If the value for field is currently set to default, desc::toggle will set it to the non-default value and vice-versa


Algorithm
Speedup
< 1
>= 1
Library
Dataset
BFS
PR
SSSP
TC
CuSha
Galois
Gunrock
Hardwired
Ligra
Mapgraph
SuiteSparse
h09
i04
rgg
rmat22
rmat23
rmat24
road_usa
soc-lj
soc-ork
cit-pat
coauthor
com-ork
copaper
road_cent
0.1
1
10
100
Value

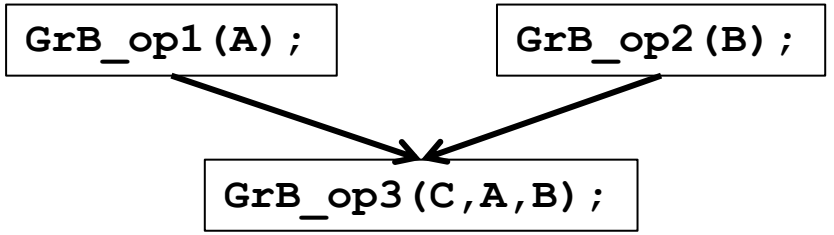
The People

- Carl Yang, who was running for the PhD position, did all the work
- John Owens, his campaign manager
- Myself, his bodyguard

GraphBLAS Execution modes

- A GraphBLAS program defines a DAG of operations.
- Objects are defined by the sequence of GraphBLAS method calls, but the value of the object is not assured until a GraphBLAS method queries its state.
- This gives an implementation flexibility to optimize the execution (fusing methods, replacing method sequences by more efficient ones, etc.)

```
GrB_op1(A);
GrB_op2(B);
GrB_op3(C,A,B);
```

- An execution of a GraphBLAS program defines a context for the library.
- The execution runs in one of two modes:
  - **Blocking mode** … executes methods in program order with each method completing before the next is called
  - **Non-Blocking mode** … methods launched in order. Complete in any order consistent with the DAG. Objects do not exit in fully defined state until queried.
- Most implementations only support Blocking mode. SuiteSparse:GraphBLAS uses nonblocking for assign and setElement

Opportunities in non-blocking mode

- Suppose you are solving a linear system on the Kronecker product graph
- Actually happens when you are computing similarity between two graphs
- Using “graph kernels” enable machine learning on graph structures data, such as proteins and other molecules.

$$\underset{N\,x\,M}{\mathbf{C}} = \mathbf{A} \otimes \underset{K\,x\,L}{\mathbf{B}} \doteq \underset{N*K\,x\,M*L}{\begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & \dots & a_{2,m}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \dots & a_{n,m}\mathbf{B} \end{pmatrix}}$$

- The Kronecker product itself has huge memory footprint and lots of redundancy (NK+ML dimension but NKML apparent values)

Opportunities in non-blocking mode

- The only way to write this in GraphBLAS or any other library we know of:

```
GrB_kronecker(C, …, A, B, …); // C=A⊗B
GrB_mxv(y, …, C, x, …); // y=C x
```

- What we would rather call:

```
GrB_kronxv(y, …, A, B, x …); // y= (A⊗B) x
```

- But that would result in API bloat and would lead us to a rabbit hole.
- There are many other examples:
  - KFAC (optimization method for deep learning),
  - Triple matrix product (graph contraction and AMG restriction),
  - Triangle counting (who needs the list of triangles when all we need is the count)

- **Solution**: **A JIT that performs automatic operator fusion**

Acknowledgments

Ariful Azad, Tim Davis, John Gilbert, Jeremy Kepner, Tim Mattson, Scott McMillan, Jose Moreira, John Owens, Oguz Selvitopi, Yu-Hang Tang, Carl Yang.

- The GraphBLAS Forum: http://graphblas.org
- My Research Team: http://passion.lbl.gov
- Graphs: Architectures, Programming, and Learning (GrAPL @IPDPS): http://hpc.pnl.gov/grapl/