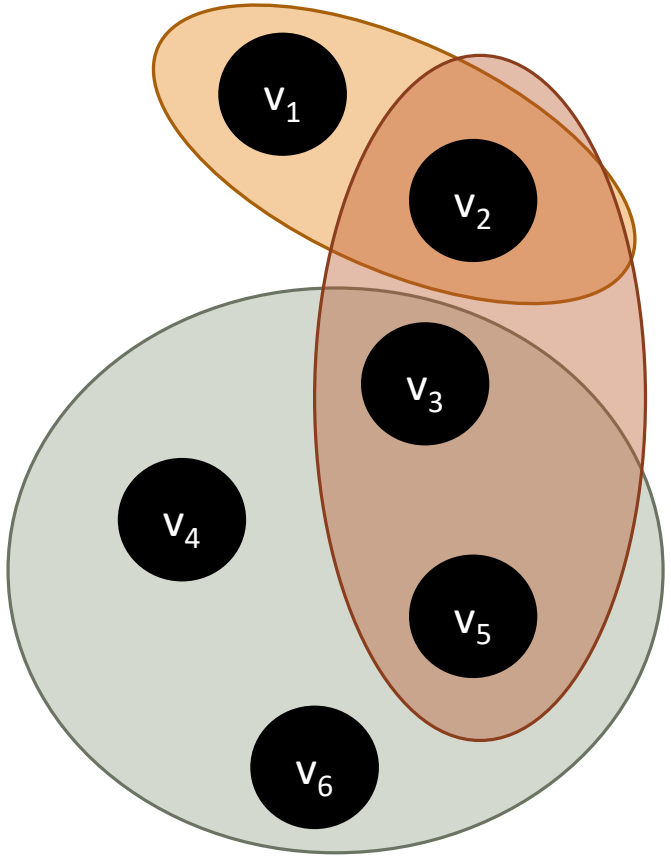
Hypergraphs in Chaos

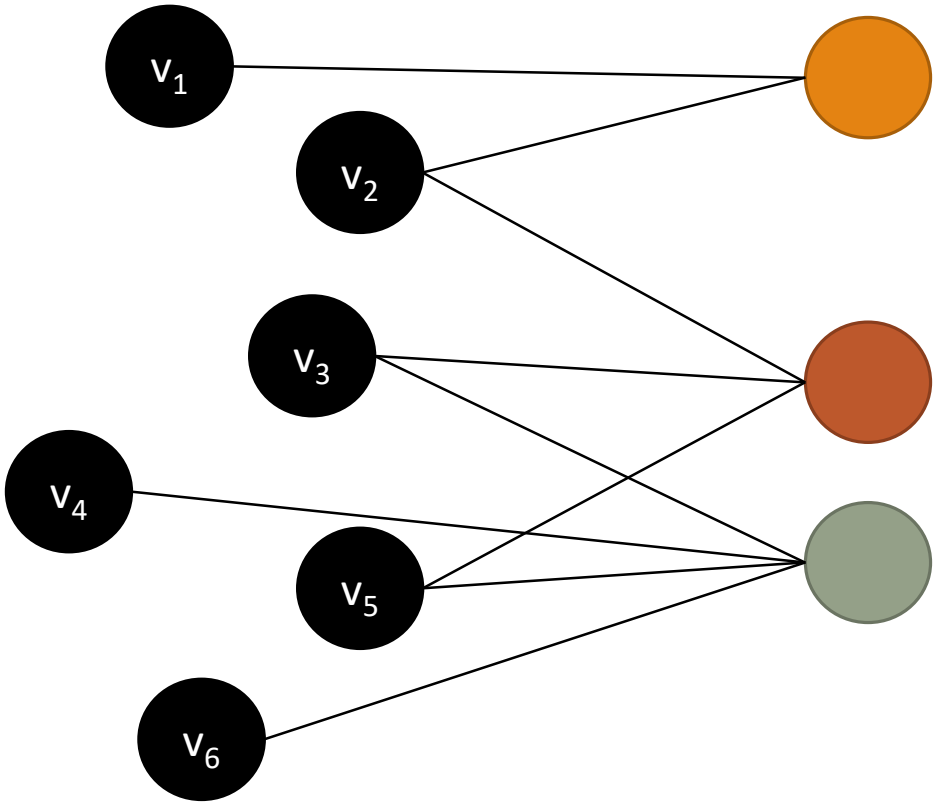
JULIUS LISCHEID

Graphs and Hypergraphs

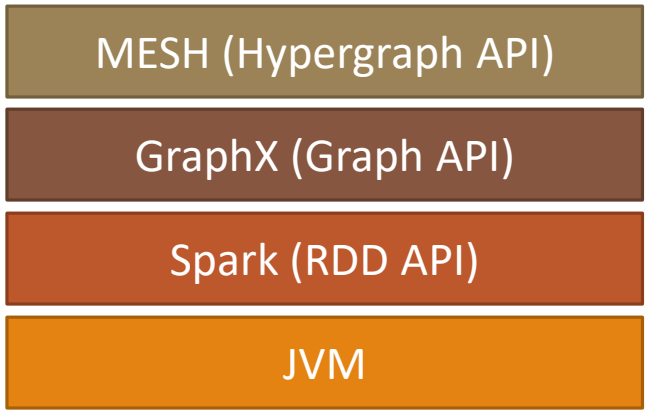


- Hypergraphs $\mathcal{H}(V, E)$ are generalised graphs where hyperedges $\mathrm{e} \in \mathrm{E}$ contain an arbitrary number of vertices $\mathrm{v} \in V$
- In short, $\mathrm{E} \subseteq \mathcal{P}(V)$
- Applications in recommender systems, image retrieval, data profiling, bioinformatics etc.

Graphs and Hypergraphs

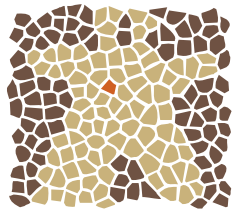


- Hypergraphs can be represented as bipartite graphs
- MESH [4], the currently fastest distributed framework, builds on GraphX that builds on Spark that builds on JVM



Distributed (Hyper)Graph Processing Genealogy

$<$ slower
$\leq$ slower or equal



(JVM) [1]

$<$

PowerGraph
(C++) [2]

$\leq$

Spark GraphX
(Spark on JVM) [3]

$\leq$

HyperX
(Spark on JVM) [5]

$\leq$

MESH: Minnesota Engine for Scalable Evolving Hypergraph Analysis
(GraphX on Spark on JVM) [4]

$<$

CHAOS
(C++) [6]

?

PowerGraph vs. GraphX

PowerGraph (C++) [2] ≤ Apache Spark GraphX (Spark on JVM) [3] ?

"[…] for graph algorithms, GraphX is over an order of magnitude faster than the base dataflow system [i.e. Spark] and is comparable to or faster than specialized graph processing systems [i.e. PowerGraph]."

Gonzalez et al., GraphX: Graph Processing in a Distributed Dataflow Framework [3]

Graph	PowerG.	GraphX	PowerL.	Gemini	Speedup (×times)
PR					
enwiki-2013	9.05	30.4	7.27	0.484	15.0
twitter-2010	40.3	216	26.9	3.02	8.91
uk-2007-05	64.9	416	58.9	1.48	39.8
weibo-2013	117	-	100	8.86	11.3
clueweb-12	-	-	-	31.1	n/a
CC					
enwiki-2013	4.61	16.5	5.02	0.237	19.5
twitter-2010	29.1	104	22.0	1.22	18.0
uk-2007-05	72.1	-	63.4	1.76	36.0
weibo-2013	56.5	-	58.6	2.62	21.6
clueweb-12	-	-	-	25.7	n/a
SSSP					
enwiki-2013	16.5	151	17.1	0.514	32.1
twitter-2010	12.5	108	10.8	1.15	9.39
uk-2007-05	117	-	143	3.45	33.9
weibo-2013	63.2	-	60.6	4.24	14.3
clueweb-12	-	-	-	56.9	n/a
GEOMEAN					19.1

Table 4: 8-node runtime (in seconds) and improvement of Gemini over the best of other systems. "-" indicates failed execution.

[7]

Project Study

CHAOS
(C++) [6]

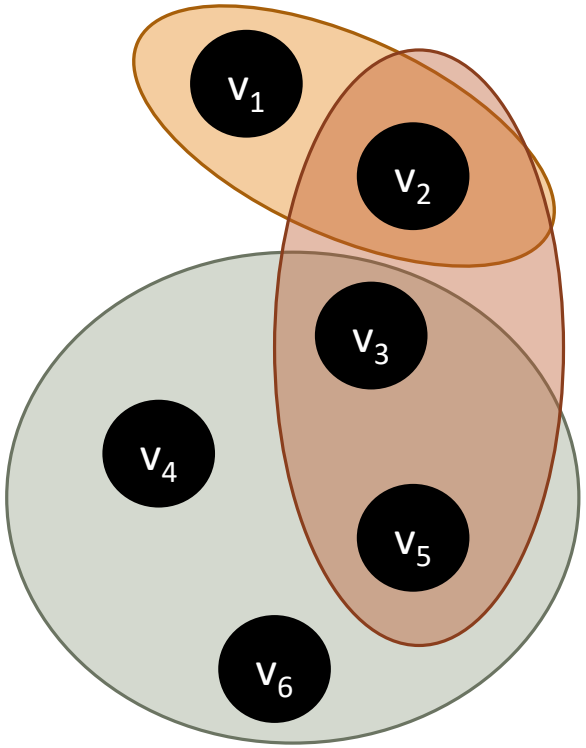
vs.

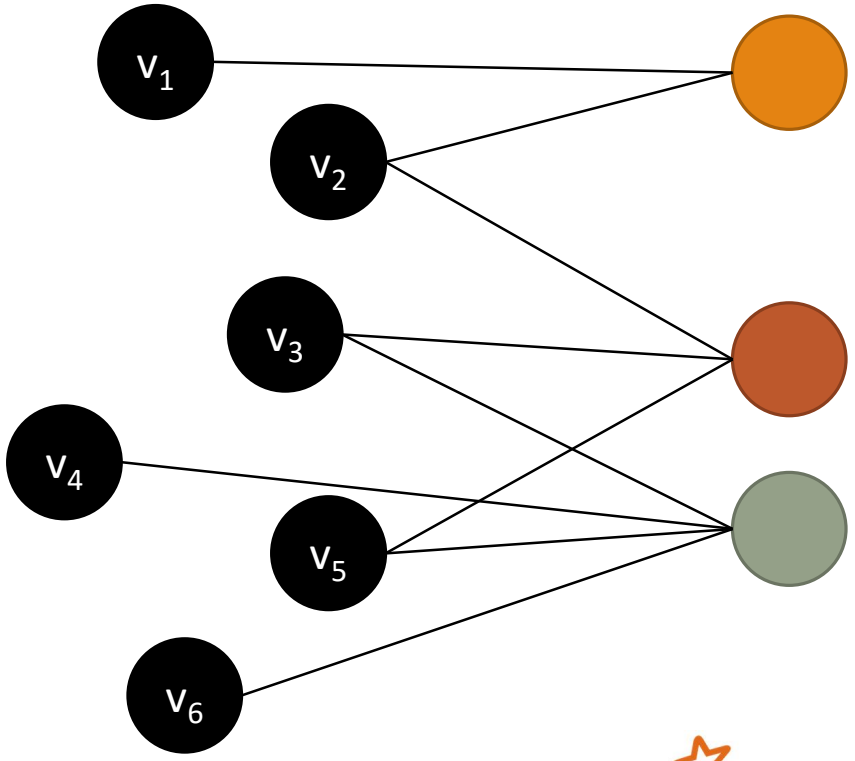
MESH: Minnesota Engine for Scalable Evolving Hypergraph Analysis
(GraphX on Spark on JVM) [4]

- Implement hypergraph PageRank algorithm in Chaos
- Benchmark it against MESH

Status Quo

```
root@julius-Virtual-Machine:/usr/local/chaos# make
g++ -O3 -DNDEBUG -Wall -Wno-unused-function -L/usr/local/lib -Wfatal-errors -DCOMPACT_GRAPH -DZLIB_COMPRESSION_LEVEL=Z_BES
T_SPEED -msse4.2 -o bin/benchmark_driver object_files/driver.o object_files/core.o object_files/utils.o -lboost_system -lb
oost_program_options -lboost_thread -lpthread -lz -lrt -lzmq -llapack
object_files/driver.o: In function `boost::property_tree::basic_ptree<std::string, std::string, std::less<std::string> >::
get_child(boost::property_tree::string_path<std::string, boost::property_tree::id_translator<std::string> > const&) [clone
 .constprop.2544]':
driver.cpp:(.text+0xfa13): undefined reference to `pt_slipstore'
object_files/driver.o: In function `slipstore::init(slipstore::io*, unsigned long, unsigned long)':
driver.cpp:(.text+0x1047b): undefined reference to `boost::program_options::abstract_variables_map::operator[](std::string
 const&) const'
driver.cpp:(.text+0x10544): undefined reference to `boost::program_options::abstract_variables_map::operator[](std::string
 const&) const'
driver.cpp:(.text+0x105d7): undefined reference to `boost::program_options::abstract_variables_map::operator[](std::string
 const&) const'
object_files/driver.o: In function `boost::program_options::typed_value<std::string, char>::xparse(boost::any&, std::vecto
r<std::string, std::allocator<std::string> > const&) const':
driver.cpp:(.text._ZNK5boost15program_options11typed_valueISscE6xparseERNS_3anyERKSt6vectorISsSaISsEE[_ZNK5boost15program_
options11typed_valueISscE6xparseERNS_3anyERKSt6vectorISsSaISsEE]+0x19): undefined reference to `boost::program_options::va
lidate(boost::any&, std::vector<std::string, std::allocator<std::string> > const&, std::string*, int)'
object_files/driver.o: In function `boost::program_options::validation_error::validation_error(boost::program_options::val
idation_error::kind_t, std::string const&, std::string const&, int)':
driver.cpp:(.text._ZN5boost15program_options16validation_errorC2ENS1_6kind_tERKSsS4_i[_ZN5boost15program_options16validati
on_errorC5ENS1_6kind_tERKSsS4_i]+0x25): undefined reference to `boost::program_options::validation_error::get_template(boo
st::program_options::validation_error::kind_t)'
driver.cpp:(.text._ZN5boost15program_options16validation_errorC2ENS1_6kind_tERKSsS4_i[_ZN5boost15program_options16validati
on_errorC5ENS1_6kind_tERKSsS4_i]+0x3d): undefined reference to `boost::program_options::error_with_option_name::error_with
_option_name(std::string const&, std::string const&, std::string const&, int)'
object_files/driver.o: In function `x_lib::configuration::setup_mapping()':
driver.cpp:(.text._ZN5x_lib13configuration13setup_mappingEv[_ZN5x_lib13configuration13setup_mappingEv]+0xf7): undefined re
ference to `boost::program_options::abstract_variables_map::operator[](std::string const&) const'
object_files/driver.o: In function `x_lib::configuration::manual()':
driver.cpp:(.text._ZN5x_lib13configuration6manualEv[_ZN5x_lib13configuration6manualEv]+0x2d): undefined reference to `boos
t::program_options::abstract_variables_map::operator[](std::string const&) const'
driver.cpp:(.text._ZN5x_lib13configuration6manualEv[_ZN5x_lib13configuration6manualEv]+0x6e): undefined reference to `boos
t::program_options::abstract_variables_map::operator[](std::string const&) const'
```





Questions?

< PowerGraph ≤ Spark GraphX (Spark on JVM) ≤ HyperX ≤ MESH: Minnesota Engine for Scalable Evolving Hypergraph Analysis (GraphX on Spark on JVM)

< APACHE GIRAPH (JVM)

< CHAOS (C++)

? (GraphX on Spark on JVM → CHAOS (C++))

References

[1] Apache Giraph. https://giraph.apache.org/

[2] Gonzalez, Joseph E., et al. "Powergraph: Distributed graph-parallel computation on natural graphs." *Presented as part of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*. 2012.

[3] Gonzalez, Joseph E., et al. "Graphx: Graph processing in a distributed dataflow framework." *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*. 2014.

[4] Heintz, Benjamin, et al. "Mesh: A flexible distributed hypergraph processing system." *arXiv preprint arXiv:1904.00549* (2019).

[5] Jiang, Wenkai, et al. "HyperX: A Scalable Hypergraph Framework." *IEEE Transactions on Knowledge and Data Engineering* 31.5 (2018): 909-922.

[6] Roy, Amitabha, et al. "Chaos: Scale-out graph processing from secondary storage." *Proceedings of the 25th Symposium on Operating Systems Principles*. ACM, 2015.

[7] Zhu, Xiaowei, et al. "Gemini: A computation-centric distributed graph processing system." *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*. 2016.