Deloitte.

Comparing Spark GraphX and Cray Graph Engine using large-scale client data

Eric Dull and Brian Sacash
Deloitte Advisory, May 11, 2017

Topics

- Introductions
- Our collection and analysis environment
- Problem statement:
  - Experimental design
  - GraphX
  - CGE
- Weaknesses and “gotchas”
- Next steps

Introductions

Deloitte Advisory's Cyber Reconnaissance team

Deloitte Advisory's Cyber Reconnaissance team uses a combination of big data tools, data science, graph analytics, and supercomputing to uncover potential threat vectors or ongoing attacks

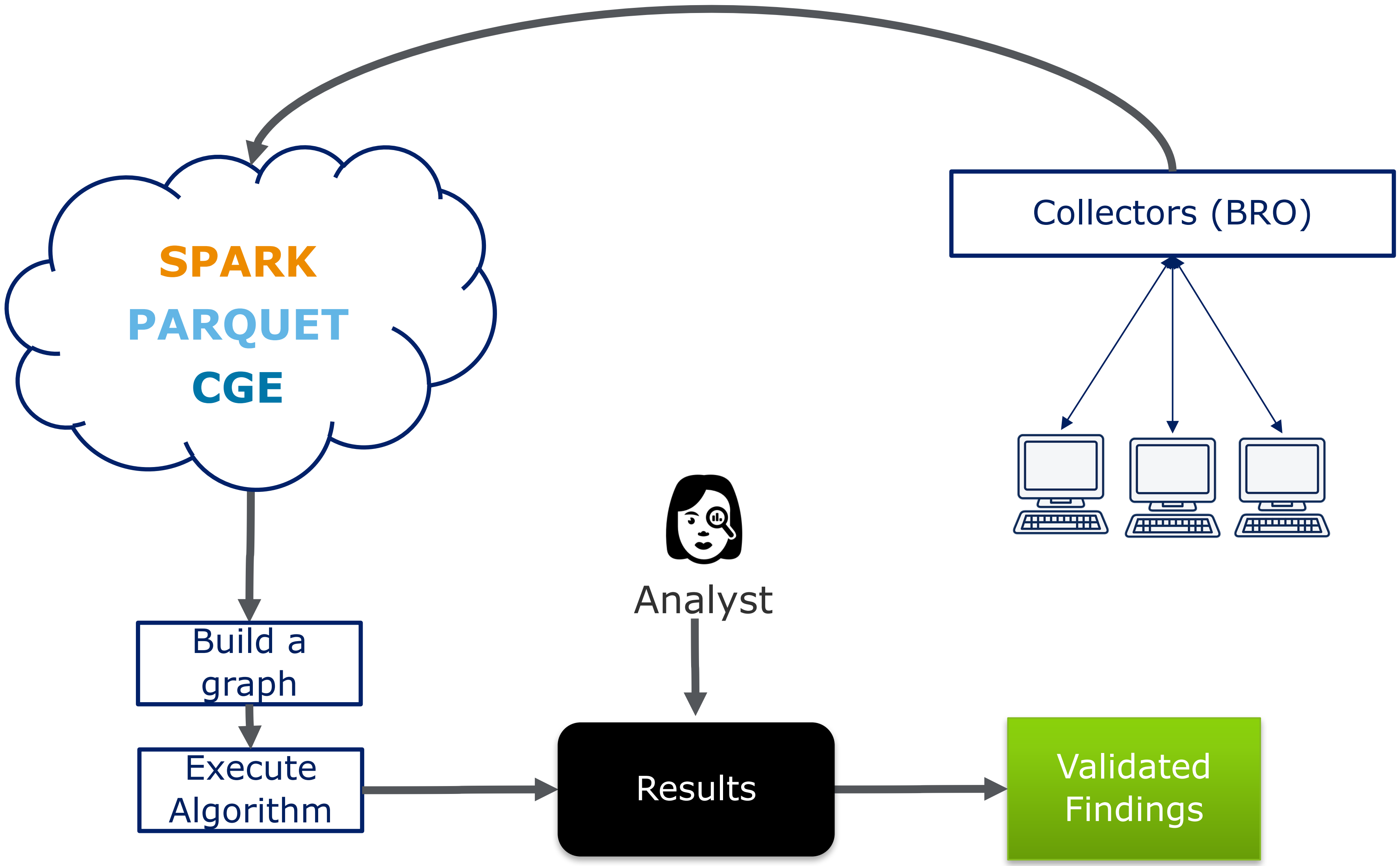
Eric Dull
Specialist Leader
Deloitte & Touche LLP

- Experience includes network analysis, applied graph analysis, behavior-based anomaly detection
- Prior CUG papers:
  - Cyberthreat analytics using graph analysis, CUG 2015

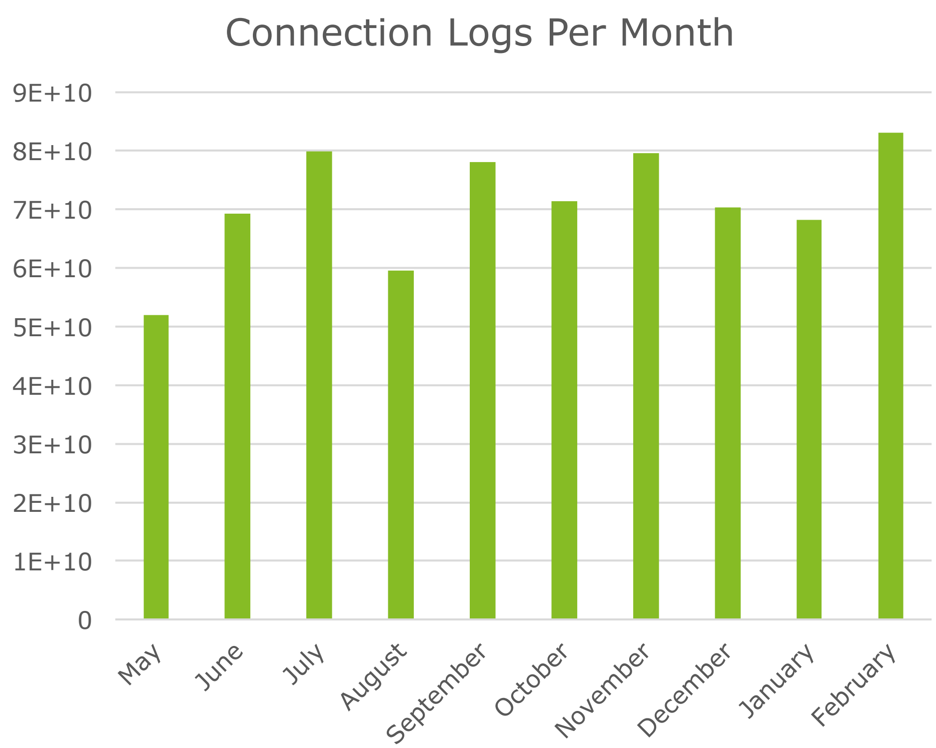
Brian Sacash
Specialist Senior
Deloitte & Touche LLP

- Data scientist who focuses on software development for analytic-based decision making
- Experience employing natural language processing, statistical analysis, and machine learning using big data technologies

Deloitte's collection and analysis environment



Data size



Source: Deloitte February data pull

February Data:

- ~83,110,000,000 connection records
- ~1,800,000 unique clients
- ~55,000,000 unique external IPs

Cray Urika-GX

What compute did we use for the experiments

Specifications

- 32 Blades
- 1000 cores
- 8 Terabytes of Ram
- 120 TB of Lustre
- 25 Blades available for Apache Spark

Additional details:

- Hosted in Deloitte's Federal Technology Center in Suwanee, GA
- Used for multiple Spark work streams supporting multiple clients

Motivation

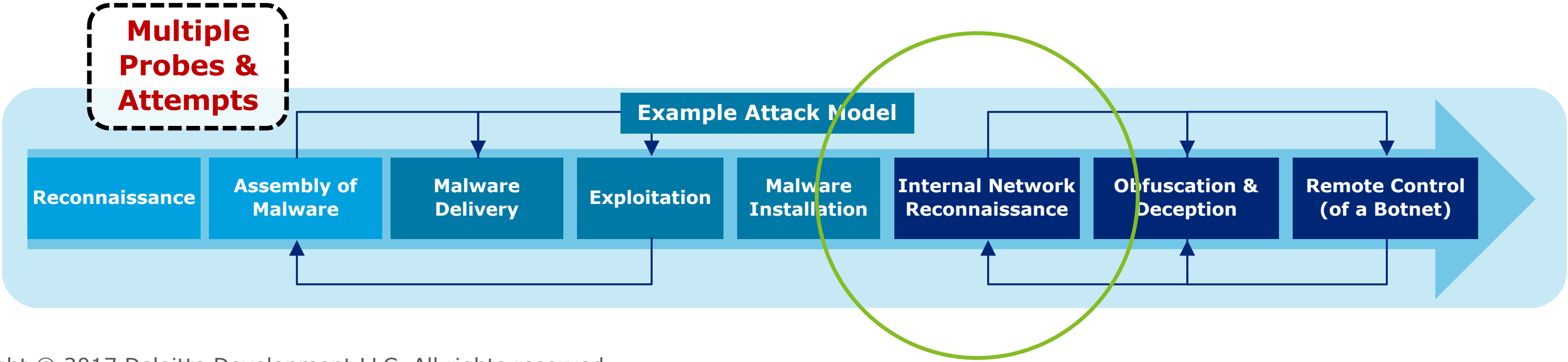
Connecting Cyber Kill Chain to Graph Algorithms

Cyber Kill Chain:

- External Reconnaissance
- Infection
- **Lurking**
- Activity

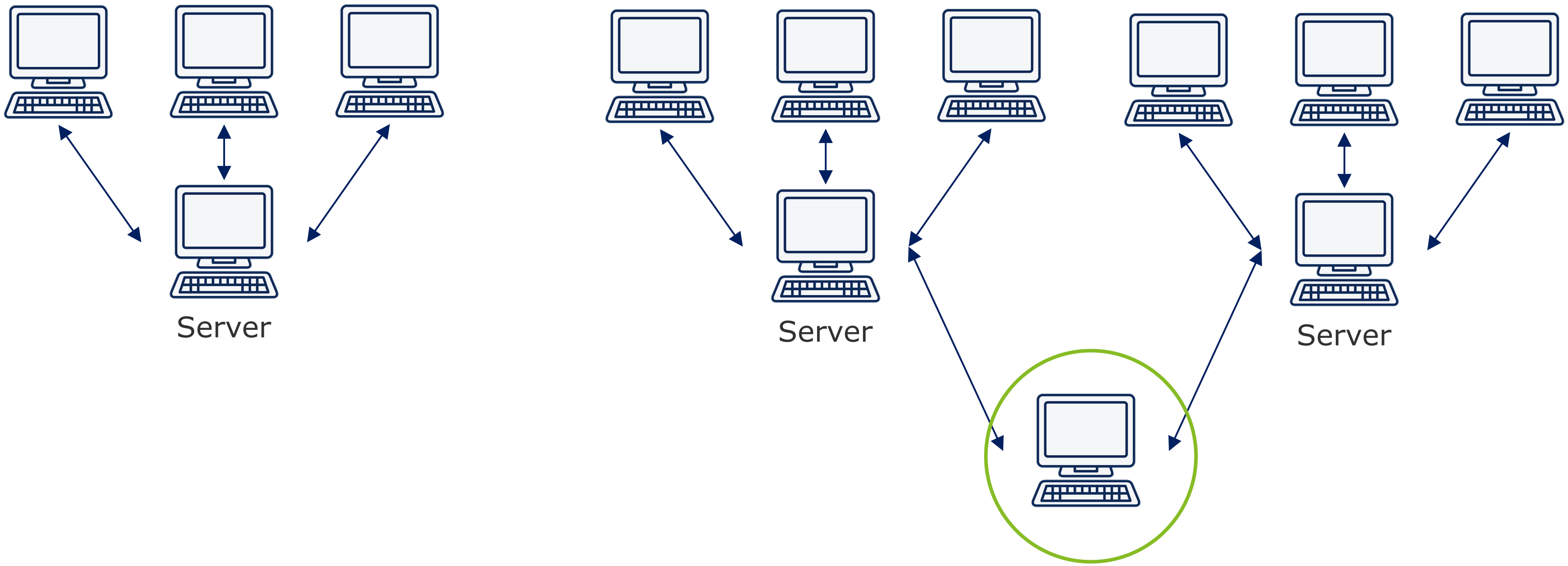
Applicable Graph algorithms:

- Community of Interest Identification
- **Betweenness Centrality**



Model

Target Graph Topologies



Betweenness Centrality: Which graph node has the most paths go through it?

Or, "All roads lead to Rome"

Experiment description

How did we execution on this vision?

Build a graph:

- Use network connection logs
- Focus on known behaviors

Run Betweenness Centrality:

- GraphX implementation
- Cray Graph Engine implementation

Validate Algorithm Results:

- Look for known scanners
- Analyst feedback

Graph Building

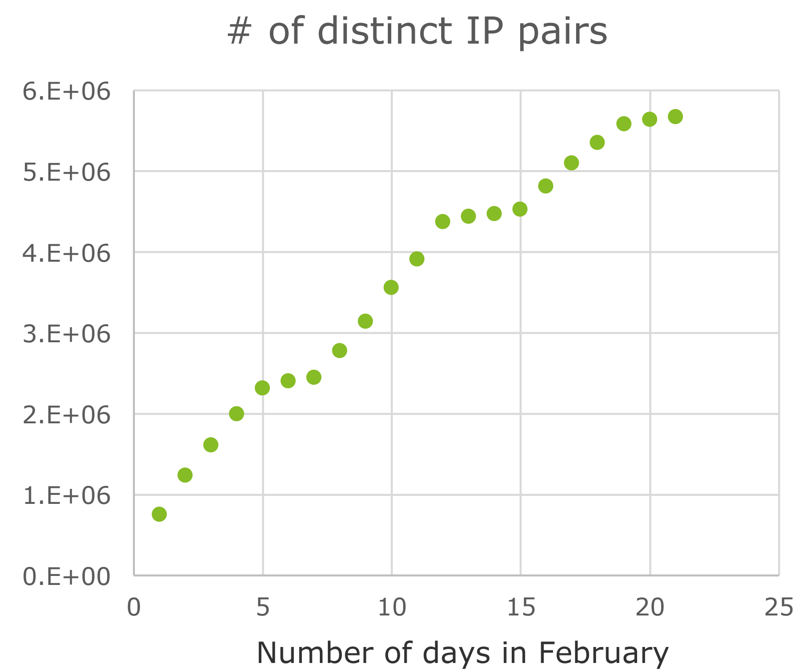
How did we build the graph?

Approach:

- Focus on TCP/UDP ports targeted by attackers
- Focus on successful connections
- Bring in multiple days

Observations:

- Successful connections reduces connection volumes by ~ 47%
- Targeted TCP ports (20, 21, 22, 23, 123, 445, 3389)
- Days remained in flux



GraphX results

How did GraphX perform?

Algorithm:

- No out-of-the-box implementation
- Spent time getting available 3rd party implementation running

Observations:

- GraphX did not perform above small graphs
- Observed variation in execution times likely related to network latency

Vertices	Edges	1 Node	4 Nodes	8 Nodes
5,419	11,726	54 seconds	50.1 seconds	71.8 seconds
42,687	125,564	N/A	N/A	N/A

CGE results

How did the CGE implementation perform?

Algorithm:

- Cray Graph Engine provides betweenness centrality callable through Sparql
- CGE implementation uses directed edges, and traditional betweenness centrality is undirected

Observations:

- CGE ran, took longer than expected
- Performance did not scale well when given additional nodes

Vertices	Edges	1 Node	8 Nodes	16 Nodes
5,419	11,726	2.6 seconds	29.4 seconds	29.3 seconds
15,359	52,042	77.7 seconds	653 seconds	718 seconds
42,687	125,564	354 seconds	1643 seconds	1891 seconds
66,955	369,720	938 seconds	4730 seconds	6600 seconds
115,276	1,281,918	3205 seconds	15068 seconds	N/A

Observations, Weaknesses, and "gotchas"

- Analyst validation in progress
- Building meaningful graphs at scale is difficult
- CGE is easy to use, and some algorithms are in progress
- GraphX is hard to use

Next steps

- Further analyst validation
- Additional algorithms / use cases
- Hybrid architectures and workflows

This document contains general information only and Deloitte Advisory is not, by means of this document, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This document is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor.

Deloitte Advisory shall not be responsible for any loss sustained by any person who relies on this document.

As used in this document, "Deloitte Advisory" means Deloitte & Touche LLP, which provides audit and enterprise risk services; Deloitte Financial Advisory Services LLP, which provides forensic, dispute, and other consulting services; and its affiliate, Deloitte Transactions and Business Analytics LLP, which provides a wide range of advisory and analytics services. Deloitte Transactions and Business Analytics LLP is not a certified public accounting firm. These entities are separate subsidiaries of Deloitte LLP. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.