



Anomaly Detection on the Blockchain

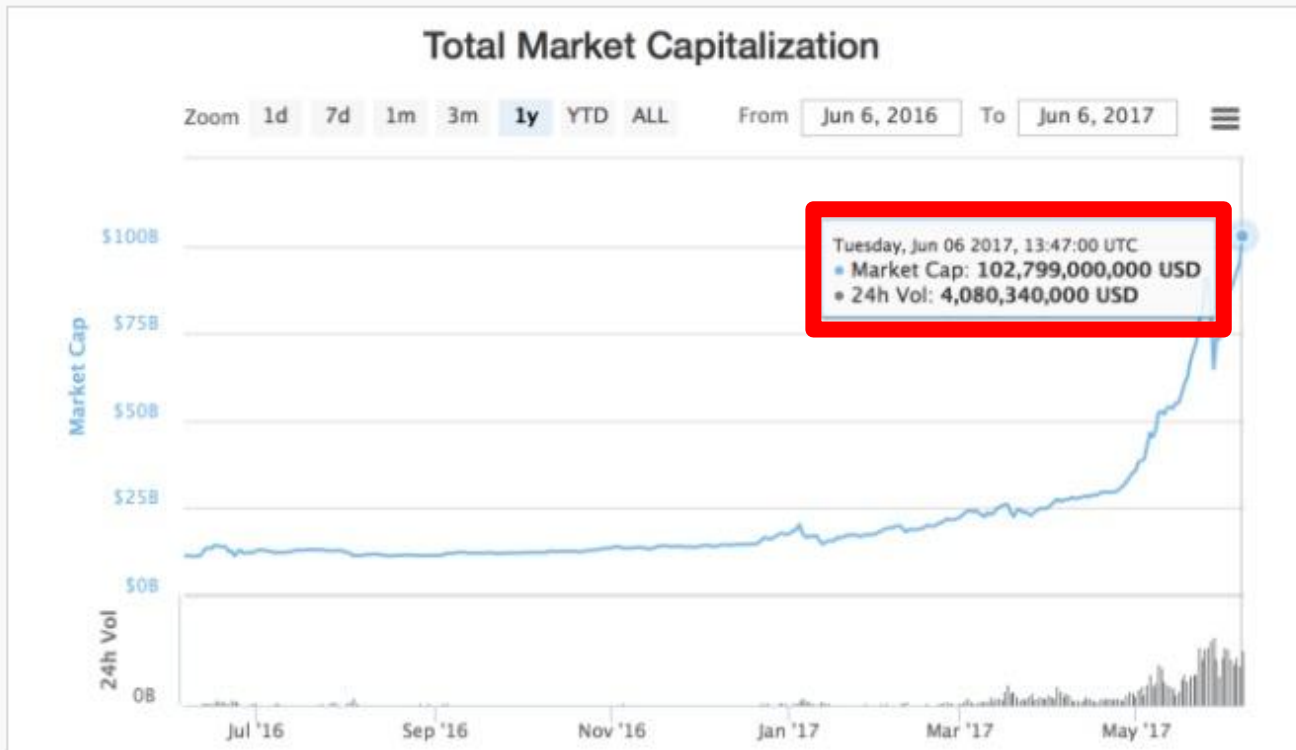
Andrew Tom

andrewtom.careers@gmail.com

Metis Data Science | June 29, 2017

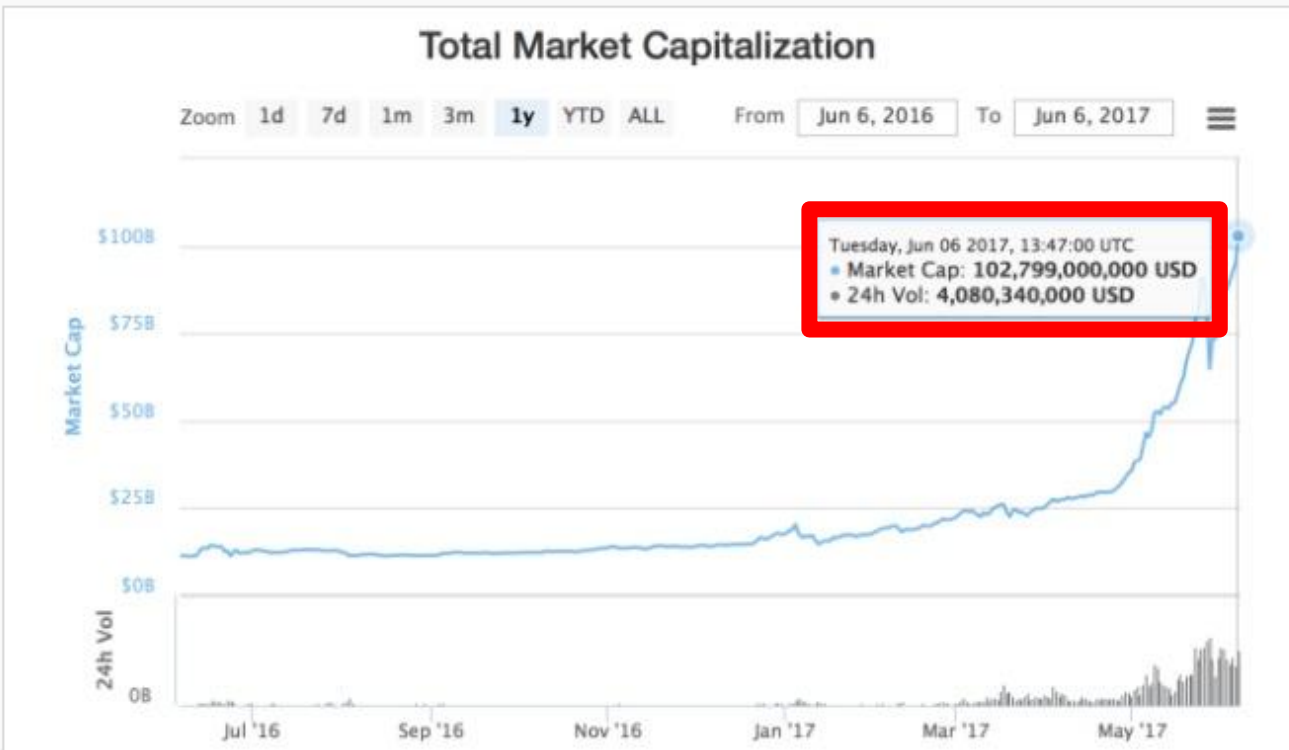


As of June 7, 2017...



Source: coinmarketcap.com

As of June 7, 2017...



Source: coinmarketcap.com



**What's holding up the
blockchain?**

Security Issues

The blockchain
needs to be secure.

Bitcoin Exchange Offers \$3.5 Million Reward for Information of Stolen Bitcoins

📅 Saturday, August 13, 2016 👤 Mohit Kumar

👍 29

684 188 43 941

\$3.5 Million Reward



For Information of \$72 Million Bitcoin Heist

Hong Kong-based Bitcoin exchange 'Bitfinex' that [lost around \\$72 Million](#) worth of its customers' Bitcoins last week is now offering a reward of \$3.5 Million to anyone who can provide information that leads to the recovery of the stolen Bitcoins.

Some examples of Bitcoin theft and fraud...

- Scam artists can front a coin exchange,
- Employees with admin access to servers can siphon funds,
- Hackers can find backdoors and penetrate exchange servers,
- Malware / trojans can infect user devices,
- Other methods yet to be discovered...

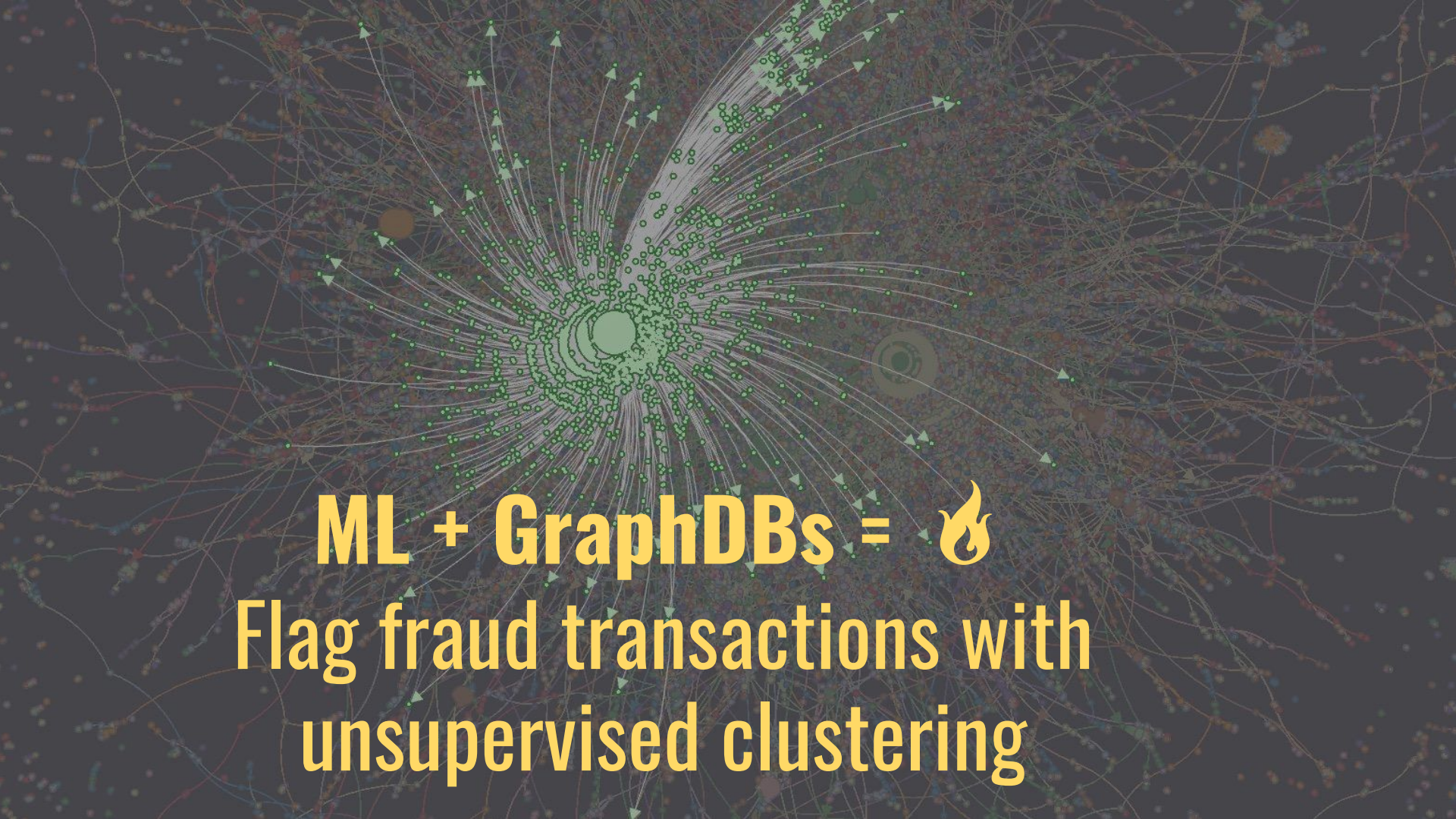
Some examples of Bitcoin theft and fraud...

- Scam artists can front a coin exchange,
- Employees with admin access to servers can siphon funds,
- Hackers can find backdoors and penetrate exchange servers,
- Malware / trojans can infect user devices,
- Other methods yet to be discovered...



How can we guard against fraud without a central authority?

What's a blockchain enthusiast to do?



ML + GraphDBs = 🔥
Flag fraud transactions with
unsupervised clustering

In a graph database...

- Instead of your typical data storage of rows and columns...
- Graph databases store data in relationships.
- How big we talking?
- The BTC Blockchain stores every transaction ever in its history!

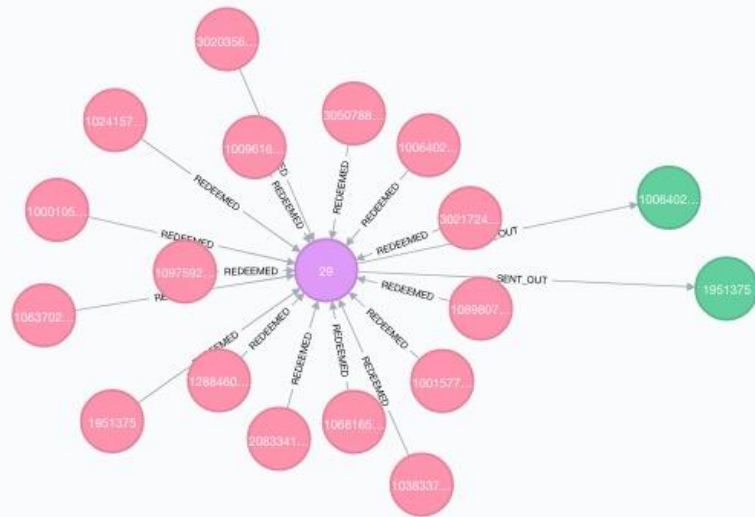


Network Stats - Neo4J + AWS

- 174M+ wallets addresses
- 1M+ annual transactions across 2.4M users since 2011

Modeled thefts from the Silk Road theft incident.

- Sep. 2013- Oct.2014.
- 543 recorded thefts
- 45,117 transactions over the same period
- Reference: <https://goo.gl/jDM6xQ>

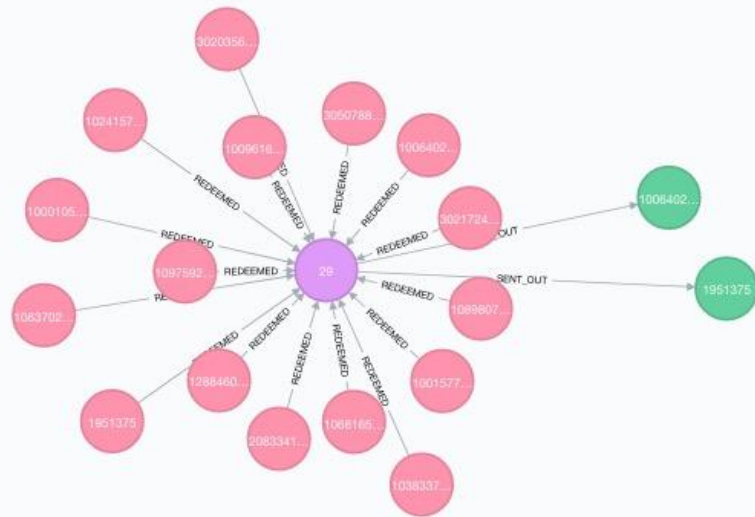


Network Stats - Neo4J + AWS

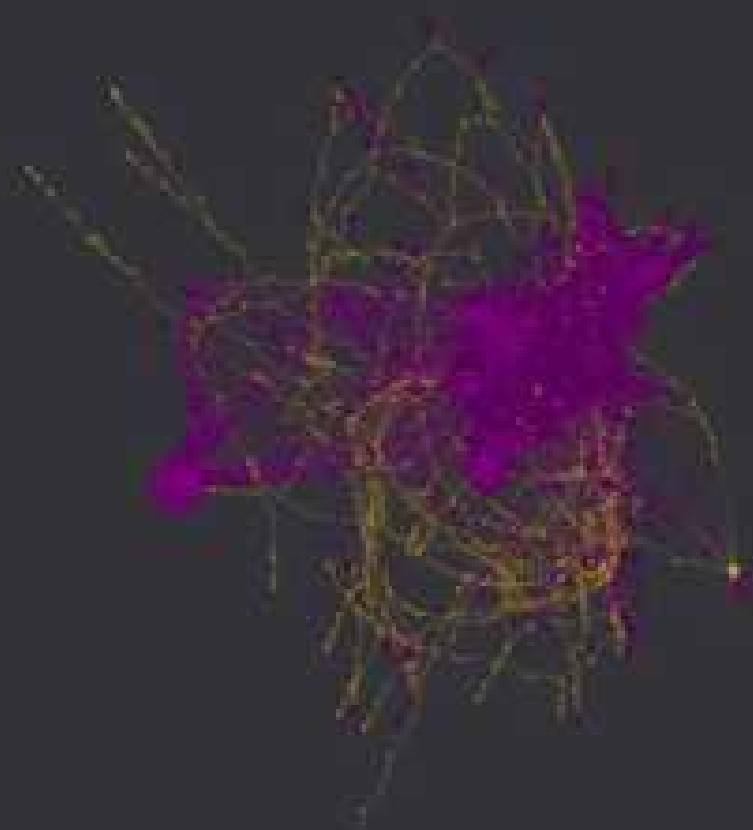
- 174M+ wallets addresses
- 1M+ annual transactions across 2.4M users since 2011

Modeled thefts from the Silk Road theft incident.

- Sep. 2013- Oct.2014.
- 543 recorded thefts
- 45,117 transactions over the same period
- Reference: <https://goo.gl/jDM6xQ>



Let's tease out unique fraud patterns using PyGraphistry



Pattern Recognition at Scale - PySpark + MLlib + Sklearn

The Model

Spark MLlib's Power Iteration

Clustering

- Normalized similarity matrix across several network statistics
- Unsupervised, Scalable, Parallelizable on Spark
- Very fast on large datasets
- 2 clusters, 10 iterations

	Precision	Recall
Thefts	100%	100%
Non-Thefts	99%	100%
homogeneity: 99%	Checks if clusters contain only members of a single class.	
completeness: 98%	Checks if all class members are assigned to the same cluster.	

Takeaways

- ML can reduce the gap between fraud detection and response.
- Unsupervised clustering is ideal for dynamic threats.
- There is potential for using graphDBs and Spark to characterize fraud at scale.
- One small step towards more secure blockchain technology.

Thank You!



Andrew Tom

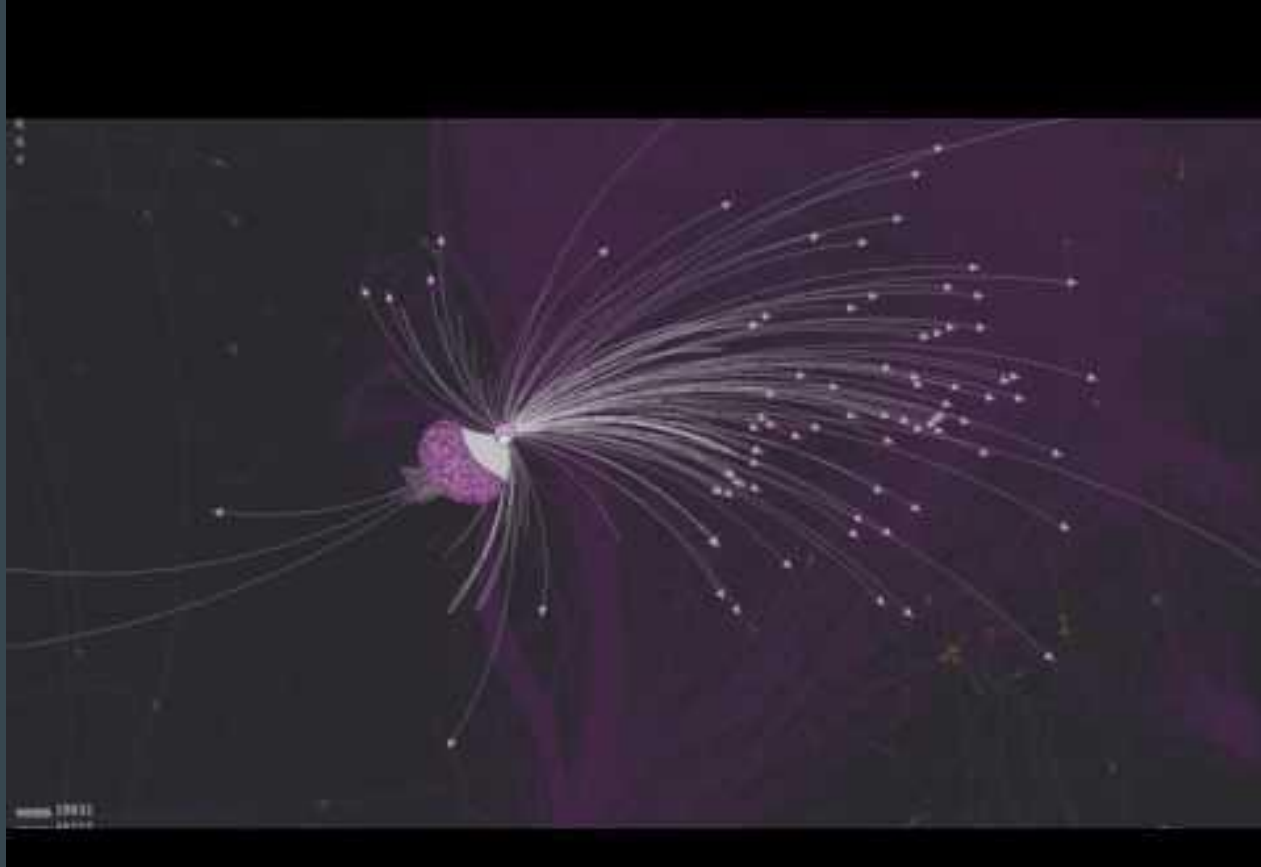
andrewtom.careers@gmail.com

Twitter: JustAndrewTom

 quarky

Appendix

Pygraphistry Music Version: <https://goo.gl/h1uHfV>



Future Work

- Couple with labeled data
- Pursue a streaming (event-driven) approach
- Assign suspicion scores and/or rules (CARTs, visual signatures) for anomalous patterns

Limitations

- Date range
- Limited to transaction patterns only, not falsified info, or identity fraud
- Class imbalance
- Batch-driven
- Lack of location
- Styles of fraud limited to labeled fraud: avg offender, criminal offender, organized crime
-
