Untangling Influence and Desire

Visual analysis of massive graph data

David Jonker, Scott Langevin

Strata + Hadoop World, London 2015
“The answer to life, the universe and everything…”

When is a simple answer too simple?
Why Seeing the Universe Matters

Believability / Trust

Perception of Truth

Pattern and Anomaly Detection

Effective visual analytics combine the best of both human and computational capabilities.
Example of a Small Universe

Sample of NBA Fan Community Chatter, Toronto Raptors
Challenges of Big Graphs as Maps

Should form a map of communities and relationships

**BUT**

- Often prohibitively costly to compute layout
- Often an unresolved hairball that reveals little
- Often prohibitively costly to render in real-time
- Ineffective labeling impedes understanding
- How to drill down from a 10,000 foot view?
Chelsea FC Fan Communities

Twitter mentions graph amongst accounts that have tweeted about Chelsea FC team or players, Aug 2014 – Feb 2015

248,747,072 tweets
554,430 account nodes
100,700 mention links
Geo Communities

Top Fans

Alex_Fitz19
Frimp_
FutbolFreeKick
BinBlueSacked
Thaneshh77
concacfoil
Stuey_95
afitzgerald218
ChelseaEnigma
RorieJamesBall3
Mitch_Rumney
benny_barrel
cthreadgold_
clinton_ijeh
dawallydahonk
Albrad5
daniireadd
charlthestrange
ayyjayy_LFC
FuckTheTile
Killingj0k3
YourBoyNaseem
HenryB_95
HodgeJoseph
MufcDevilUpdate
NariahCFC
ArabVegeta
ELLY__MARSH
ChrisMears1
NazSadikHazard
JamesGholami
jamieberry11
HowardRyan_
LukeRomeo
Lozpidicueta
matthewlynch91
the_reveilator
rossstanton_
JRowellLFC
Silterrific
joeychelsea588
jRussell_10
HodgeJoseph
GingerMatt
EvanMulligan
dpointheady
DarrenDoyleCFC
JackONeillLFC
sullivan_craig5
iRonanCooper
andy_brudz
ColayBear
DaraghCavanagh
GlennArla
HuwRyall
LazarTag
kathb24
harveyjj24
Richards___1905
TalousTheeem98
naseem95
RedFish_8
Paalmer
evanfrancisc097
samenebelli
OlyCFC56
TroubleMan19
DavidTReeves
CALSMITH
Regalinho
cbrenta
GFCOwen
Robert15puzzles7
Geo Communities

Top Hashtags
Technical Approach

**Tile-Based** Community Graph Visualization

- Distributed hierarchical community clustering.
- Hierarchical community-based graph layout.
- Multi-scale node and link tile-based rendering with LOD.
- Layered community analytics.
Aperture Tiles

TILE-BASED VISUAL ANALYTICS
• Hierarchical data tiling using cluster computing.
• Interactive on-demand image tile generation.
• Layers of raw data and derivative analytics.

OPEN SOURCE
• Uncharted research product.
• Built on Apache Spark and Hadoop.

• Cross Plots, Geospatial, Time-series, Graph analysis
Three Tiers

**Tile Generation**

- Generates tiled data views from raw source data

**Tile Service**

- Generates rasters, query responses from tiled data views

**Tile Client**

- Consumes tile rasters and data
Graph Tiling Pipeline

Source Data
GraphML or CSV

Hierarchical Community Clustering

Hierarchical Graph Layout

Tile Generation

Tiled Nodes + Links

Optional
Hierarchical Community Clustering

- Raw Graph
- Level 1 Communities
- Level 2 Communities...

Distributed Louvain Modularity

GraphX

ML

Graph ML

CSV

Hadoop

Spark

Hierarchical Communities
Hierarchical Graph Layout

Level 2 Communities Layout

Level 1 Communities Layout

Raw Nodes Layout

Hierarchical Communities

Hierarchical Force Directed Layout

Spark GraphX

Hadoop

Graph Layout
Hierarchical Tile Generation

Graph Layout → Tile Generation → Tiled Data Views

- Aggregated Nodes + Edges
- Community Analytics
Product Affinity

Stanford collection of anonymized Amazon product reviews and “customers who bought this also bought…” over 9 years

2,372,409 product and customer nodes
9,909,551 review and co-purchase links

Demo

skip
<table>
<thead>
<tr>
<th>Name</th>
<th>Entity Class</th>
<th>Popularity Percentile</th>
<th>Reviews</th>
<th>Avg Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Hitchhiker's Guide to the Galaxy</td>
<td>Book</td>
<td>100</td>
<td>579</td>
<td>4.5</td>
</tr>
<tr>
<td>The Ultimate Hitchhiker's Guide to the Galaxy</td>
<td>Book</td>
<td>100</td>
<td>73</td>
<td>4.5</td>
</tr>
<tr>
<td>The Hitchhiker's Guide to the Galaxy</td>
<td>Book</td>
<td>99</td>
<td>171</td>
<td>4.5</td>
</tr>
<tr>
<td>The Hitchhiker's Guide to the Galaxy : A Novel</td>
<td>Book</td>
<td>96</td>
<td>171</td>
<td>4.5</td>
</tr>
<tr>
<td>The Hitchhiker's Guide to the Galaxy</td>
<td>Book</td>
<td>96</td>
<td>171</td>
<td>4.5</td>
</tr>
<tr>
<td>The Hitchhiker's Guide to the Galaxy</td>
<td>Book</td>
<td>96</td>
<td>171</td>
<td>4.5</td>
</tr>
<tr>
<td>The Hitchhiker's Guide to the Galaxy</td>
<td>Book</td>
<td>96</td>
<td>171</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Global Product Affinity Graph
Bitcoin Transaction Correlations
Silk Road Bitcoin Sources within Global Communities
Other Applications

Doc Graph

Neural Networks

more to come
How much of the big picture is missed with simple charts? Combining computational analytics with highly expressive interactive visualization affords deeper understanding and trust. Map techniques and cluster computing are highly applicable.

Behavior and influence moves beyond What, to Why and How.
What’s Next?

Products

Graph Tiling Pipeline

- Enhanced analytic layers for communities.
- Integration of runtime graph analytics.
- Further optimization.