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

Fans Mention Graph
Geo Communities

Top Fans

- Alex_Fitz19
- Frimp_
- FutbolFreeKick
- BinBlueSacked
- Thaneshh77
- concalcio
- sf Fitzgerald218
- ChelseaEnigma
- Rorie_JamesBall3
- Mitch_Rumney
- benny_barrel
- ctthreadgold
- clinton_ljeh
- dawallydahong
- Airbad5
- danireadd
- charlthestrange
- ayyjayy_LFC
- FuckTheTitle
- Killingj0k3
- YourBoyNaseem
- HenryB_95
- HodgeJoseph
- MucoDevilUpdate
- NariahCFC
- ArabVegeta
- ELLY__MARSH
- ChrisMears1
- NazSadikHazard
- JamesGholami
- jamieberry11
- HowardRyan_
- LukeRomeo
- Lozpicueta
- Matthewlynch91
- the_revelator
- rossstanton
- JRowellLFC
- Salterrific
- joeychelsea588
- jRussell_10
- HodgeJoseph
- GingerMatt
- EvanMulligan
- djoinehealy
- DarrenDoyleCFC
- JackONeillLFC
- sullivan_craig5
- iRonanCooper
- andy_brudz
- ColasBear
- DarraghCavanagh
- GlennArfa
- HuwRyall
- LazarTag
- kathb24
- harveyj24
- Richards_1905
- TalalThaneem98
- RedFish_8
- Paalmer
- naseem95
- evanfrancis97
- sàmenebeli
- OllyCFC56
- TroubleMan19
- DavidTReeves
- CALSMITH
- Regalinho
- cbrenta
- GFC Owen
- Robert15puzzles7
- Matthewlynch91
- the_revelator
- rossstanton
- JRowellLFC
- Salterrific
- joeychelsea588
- jRussell_10
- HodgeJoseph
- GingerMatt
- EvanMulligan
- djoinehealy
- DarrenDoyleCFC
- JackONeillLFC
- sullivan_craig5
- iRonanCooper
- andy_brudz
- ColasBear
- DarraghCavanagh
- GlennArfa
- HuwRyall
- LazarTag
- kathb24
- harveyj24
- Richards_1905
- TalalThaneem98
- RedFish_8
- Paalmer
- naseem95
- evanfrancis97
- sàmenebeli
- OllyCFC56
- TroubleMan19
- DavidTReeves
- CALSMITH
- Regalinho
- cbrenta
- GFC Owen
- Robert15puzzles7
Geo Communities

Top Hashtags
Community Graph

#breakfastinthecity #superdrivetimeshow #rc1023fm #citycafe #np #mufc #mellowworld #nowplaying #continentalsunrise #continentaldrive #ladelciclonenfc #futboltotaldirectv #ladelcampeonenfc #globoesperimaefc #ladernomanen1190game... #igabbva #halaamadrid #laderiverenfc #laderacing17enfc #realmadrid #namastoeobama #carregabenifica #peace #mufc #alltogethernow #ontopictalkshow #ofemerides #final #missiondelhi #footballremember #goluruguay #futboltotaldirectvrenovado #nbas #hipismo #jamasloavistastealidirectv #editorial #lvb #deportes #noticiascolorido #elmarcador
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.
  o 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

Spark

Hadoop

HBase
Hierarchical Community Clustering

Raw Graph → Level 1 Communities → Level 2 Communities → ...

Graph ML → CSV → Distributed Louvain Modularity → Spark GraphX → 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
Product Search

Showing 11 of 11 results, ordered by Entity Class, Popularity Percentile

<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>
<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>
Influential Customers

- Elijah Potino (+257)
- Scottie Baumann (+11)
- Kelly Billingsley (+483)
- Sheila Raines (+106)
- See Clement (+6)
Global Product Affinity Graph
Bitcoin Transaction Correlations
Silk Road Bitcoin Sources within Global Communities
Other Applications

Doc Graph

Neural Networks

more
to come
Summary

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.
Questions?