Overview

1. What’s a graph and why are they hard to visualize?

2. Visualizing graphs based on what we want to learn through some examples:
   - Who’s close and who’s opposite?
   - Did these people meet?
   - Who’s connected to me and my competition?
   - What’s are the characteristics of this big graph?

3. I have some graph data: now what?
   - Point and click
   - Programming and open source tools

4. What next?
Seven Bridges of Königsberg problem (1735 – Leonhard Euler)
- Find a walk through the city that crosses each bridge once and only once.
A tree is not the same as a graph. A tree branches but does not have loops. Family trees are trees, usually (except in this family tree from the house of Stuart when Mary Queen of Scots marries her cousin Henry Stuart.)
Graph 101

Characters in Les Miserables who appear with one another. There are many connections so interaction helps show who’s connected to Cosette.

from Gephi: open-source point-and-click graph visualization. Download it after this talk! www.gephi.org
Same graph of Les Miserable – but shown as a matrix.

In a matrix, nodes are listed along the side and top; and a dot occurs in the matrix for two characters that appear with one another.

The row for Cosette has been highlighted with the corresponding connections to other nodes highlighted.

From: [http://bost.ocks.org/mike/miserables/](http://bost.ocks.org/mike/miserables/)
Graph 101

LinkedIn Graph for me:
Graph 101

Apple’s board members connections to other company’s boards and their board members. Interconnections are difficult to see.

from Gephi: open-source point-and-click graph visualization. www.gephi.org
The Challenge

When graphs get big, how do you make sense of them?
Making Sense of Big Graphs #1

Who’s close?
Who’s opposite?

→ correlation graph
Visualizing a Correlation Graph

Step 1. Timeseries Correlation
Any two timeseries of data may have similar movement:

Google (GOOG) and Exxon (XOM) Stock Price (1 year)
Visualizing a Correlation Graph

Step 2. We can calculate this relationship as a correlation.

a. Plot the daily percent change as a scatterplot.
b. The linear regression shows the relationship as slope:

c. A positive slope, like this, indicates strong correlation.
Visualizing a Correlation Graph

Step 3. We can express this correlation as a single number and do this for all the items of interest.

<table>
<thead>
<tr>
<th>Ticker</th>
<th>XOM</th>
<th>GOOG</th>
<th>CVX</th>
<th>IBM</th>
<th>PG</th>
<th>AAPL</th>
<th>MSFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>XOM</td>
<td>0.892</td>
<td>0.700</td>
<td>0.639</td>
<td>0.677</td>
<td>0.268</td>
<td>0.640</td>
<td></td>
</tr>
<tr>
<td>GOOG</td>
<td>0.892</td>
<td>0.667</td>
<td>0.688</td>
<td>0.672</td>
<td>0.301</td>
<td>0.512</td>
<td></td>
</tr>
<tr>
<td>CVX</td>
<td>0.700</td>
<td>0.667</td>
<td>0.867</td>
<td>0.436</td>
<td>0.751</td>
<td>0.117</td>
<td></td>
</tr>
<tr>
<td>IBM</td>
<td>0.639</td>
<td>0.688</td>
<td>0.867</td>
<td>0.516</td>
<td>0.843</td>
<td>-0.075</td>
<td></td>
</tr>
<tr>
<td>PG</td>
<td>0.677</td>
<td>0.672</td>
<td>0.436</td>
<td>0.516</td>
<td>0.283</td>
<td>0.476</td>
<td></td>
</tr>
<tr>
<td>AAPL</td>
<td>0.268</td>
<td>0.301</td>
<td>0.751</td>
<td>0.843</td>
<td>0.283</td>
<td>-0.405</td>
<td></td>
</tr>
<tr>
<td>MSFT</td>
<td>0.640</td>
<td>0.512</td>
<td>0.117</td>
<td>-0.075</td>
<td>0.476</td>
<td>-0.405</td>
<td></td>
</tr>
</tbody>
</table>

1 indicates a strong relationship - both prices move together in unison.
0 indicates no relationship.
-1 indicates an inverse relationship.

Correlations are important to financial portfolio managers:
A diversified portfolio should have holdings that are not correlated (i.e. correlations close to zero).
A hedged portfolio may have holdings with inverse correlations (i.e. negative correlations).
Visualizing a Correlation Graph

Correlations in a matrix. Difficult to scale:

With 200 items, there are now 40,000 cells (200x200)>
> patterns such as clusters and outliers are not
easily discernable
Instead: Draw it as a node + link graph:

- 200 stocks around perimeter
- Every correlation shown as a line
- Too many lines to gain any insight
These are the strong connections...

- Remove most of the lines - only correlations between .85 - 1.0.
- Still too many lines to gain insight.
So pull together the strong connections, push everything else apart

Exxon (XOM) and Google (GOOG) are highly correlated and very close together.

This group of stocks down here are all highly correlated.
But wait – turn on .75-.85 correlation links.

Some of these items are far apart but still have a fairly strong connection.
How about putting it on a 3D sphere?

Clustering on a sphere design premise –

- For a given point:
  1. Strong correlations are close by.
  2. Negative correlations are opposite.
  3. Non-correlations are orthogonal.
Sphere Correlation

Answers real questions that financial managers are interested in e.g.:
- What else trades similar to this stock?
  - Is it all alone?
- What’s opposite to this stock?
- What can I get that has no connection to this stock?

Exxon (XOM) and Google (GOOG) are near each other

Apple (AAPL) is opposite of Microsoft (MSFT)
By the way, you can correlate other kinds of timeseries data in the same way, for example – Twitter:

Justin Bieber is on the opposite side from the Harvard Business Review (i.e. he Tweet timing pattern is opposite that of HBR’s).
Demo: [http://www.oculusinfo.com/assets/demos/SphereTreeDemo/index.html](http://www.oculusinfo.com/assets/demos/SphereTreeDemo/index.html)
Correlation Graph Summary

1. The actual links didn’t need to be visible – proximity sufficiently answered the investment questions.

2. Use a layout that is both:
   + effective (in this case proximity)
   + intuitive (sphere works well for “opposite”)
Making Sense of Big Graphs #2

Did these people meet?

→ graphs in time and space
Graph on a map

One taxi starting in San Francisco then driving around Oakland.
Two taxis starting in San Francisco then driving around Oakland. Did they meet?
Graph on a map + time in third dimension

Two taxis, again. Vertical axis represents time. Now potential crossings are more visually discernable.
Find crossings/meetings/…

Analytic “pattern finders” can identify patterns such as potential locations where many taxis may have met.
Scale up: 100,000 nodes

Many more taxis, many more points. Again, analytics can be used to find patterns.
Time and Space Graph Summary

1. Pull a dense graph apart to see what’s going on:
   + zoom/pan
   + filtering
   + 3D

2. Use graph analytics to identify patterns of interest:
   + proximity
   + number of connections
   + density of nodes
Making Sense of Big Graphs #3

follow the money?

→ flow graph
Money Flow: Charity Network

5,000 charities. 2,000,000 donors. 5,000,000 donations.

<table>
<thead>
<tr>
<th>Charities by # of Donors</th>
<th>Charities by Total Donations</th>
<th>Donors</th>
</tr>
</thead>
<tbody>
<tr>
<td>9% only 1 donor</td>
<td>29% less than $10,000</td>
<td>67%</td>
</tr>
<tr>
<td>63% 1 – 100 donors</td>
<td>38% $10,000 - $100K</td>
<td>32%</td>
</tr>
<tr>
<td>23% 100 – 1000 donors</td>
<td>26% $100K - $1M</td>
<td>1%</td>
</tr>
<tr>
<td>4% 1000 – 10000 donors</td>
<td>6% $1M - 10M</td>
<td>&lt;1%</td>
</tr>
<tr>
<td>&lt;1% 18 charities have 10000+</td>
<td>&lt;1% 30 received $10M+</td>
<td>21 gave to 100+</td>
</tr>
</tbody>
</table>

Anonymized, public, open data.

3 weeks to design and build an exploratory web-based tool.

For any charity:

Who are my donors?

Who are my prospective donors?
My Donors

1. Biggest donations to me.

2. Base of support in Pacific North-West.

3. Donations were mostly in the summer of 2012.
My Competition?

This cluster took in the greatest number of large donations... ...but mostly in the Winter of 2011 / 2012....

...and its support is mostly national, with strength in CA and the South...
My Competition

...is drawing from the same donor region...

...at the same time...

This charity...

...and has the biggest donor overlap. We want these $!
Untapped Donors?

Organic layout draws communities together

Can see some big donations here

But interconnected nodes create "hairball" problem.

Adaptive labeling tags most important of the 2000 nodes
Untapped Donors: Community Aggregation

Money flow is more clearly summarized.

Aggregation markers summarize size and makeup of communities.

© 2013 Oculus Info Inc
My competitor has a high percentage of large untapped donors. I want these too!

Left to right layout improves reading of flow direction.
Charity Network: Summary

- **Adaptive labeling** using fast label deconfliction[^2] can be used to label most important of many nodes in available display space without obscuring valuable data.

- **Clustering** and aggregation of communities, with **visual aggregation markers**, can more effectively communicate big graphs without information loss.

[^1]: Louvain Clustering for Big Data Graph Visual Analytics (Gauldie, 2013)

[^2]: Fast Point-Feature Label Placement for Dynamic Visualizations (Mote, 2007)

[^3]: Fast unfolding of communities in large networks (Blondel, 2008)
Money Flow II: Influent

3 months to design and build a tool for financial forensics:
- Use similar clustering and aggregation techniques for scalability.
AND
- See temporal and geopolitical patterns across clusters simultaneously.
- “Follow the money” through any number of steps.
- Interactive drill down into interesting clusters.
- Easily integrated with, and tailored for, different data sets.

Test using public, open data sets:

- Kiva
  - 2,000,000 accounts. 4,000,000 transactions.

- Bitcoin
  - 6,000,000 accounts. 16,000,000 transactions.
Example: Kiva Partner Transactions

Began by searching for a partner of interest. Found and filed a match.

Icons summarize cluster characteristics such as country of origin.

Clicked [+] buttons to follow money left to right. Incoming from microloan lenders, outgoing to borrowers.
Example: Kiva Partner Transactions

Histograms summarize payments over time, highlighting transactions involving entity of interest.

Unstack clusters by clicking paper clips to drill down.
Example II: **Bitcoin Transactions**

**Example Question:** Where do WikiLeaks Donation Funds Originate?
Example II: Bitcoin Transactions

...and where do they go?
Example: Bitcoin Transactions

Account is emptied of almost $100,000 one day later.

Six transactions to this account

Equivalent amount moves through accounts on same day.
Example: Bitcoin Transactions

Equivalent amount moves through accounts on same day.

x 30
Example: Bitcoin Transactions

+similar account ids. why?
Money Flow Graph Summary

1. Focus on the flow
   + search to start
   + expand

2. Analytics to handle complexity e.g.
   + clustering
   + similarity
Making Sense of Big Graphs #4

What’s in this graph?

→ any graph
e.g. every bitcoin transaction
Bitcoin Transactions
Sender (x) by Receiver (y)

7 zoom levels
4 billion pixels
Aperture Tile Based Visual Analytics

Technology stack for scalable, zoomable 2D and 1D plots.

Using:

Spark

Apache HBase

ApertureJS™ [aperturejs.com]

Acknowledgements: These studies were supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Number FA8750-12-C-0317. The views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.
Matrix Graph Summary

Use a matrix graph for exploratory analysis:
1. does it look right:
   + gaps/errors
2. look for patterns:
   + horizontal/vertical lines
   + diagonal lines
   + bright clusters
Many more techniques exist

Edge Bundling

Hive Plot

Hierarchical/Euler

Sankey Diagrams

Circos

http://bost.ocks.org/mike/hive/
http://egweb.bcgsc.ca/

http://www.csse.monash.edu.au/~tdwyer/

http://bost.ocks.org/mike/sankey/

http://circos.ca/
Where next?

Want to point and click through some graphs?
- Gephi (for node and link graphs)  [www.gephi.org](http://www.gephi.org)
- Cytoscape (another for node and link graphs)  [www.cytoscape.org/](http://www.cytoscape.org/)
- Excel (for matrix graphs)

Want to program using some toolkits?
- D3.js  [d3js.org](http://d3js.org)

Coming soon to [github.com/oculusinfo](http://github.com/oculusinfo):
Some nice reference books

See some really nice looking graph **pictures** for inspiration?

→ **Visual Complexity:**
   **Mapping Patterns of Information**
   by Manuel Lima

See some **code** for drawing nice looking graphs so I can start programming!

→ **Graph Drawing:**
   **Algorithms for the Visualization of Graphs**
   Ioannis G. Tollis (Author), Giuseppe Di Battista (Author), Peter Eades (Author), Roberto Tamassia (Author)
Summary

Key takeaways
- Graphs are challenging to perceive patterns when big
- Start with your analytic question – what are you trying to find
  - Format: e.g. node and link, 3D, flow, matrix
  - Interactions: e.g. pan/zoom, filter, etc
  - Analytics: clustering, connections
- Many possibilities and variants

More info
- Richard Brath
  richard <dot> brath (at) oculusinfo <dot> com
- David Jonker
  david <dot> jonker (at) oculusinfo <dot> com