Big Graphs and Social Network Visualization

Strata NYC 2013

Richard Brath
David Jonker
www.oculusinfo.com

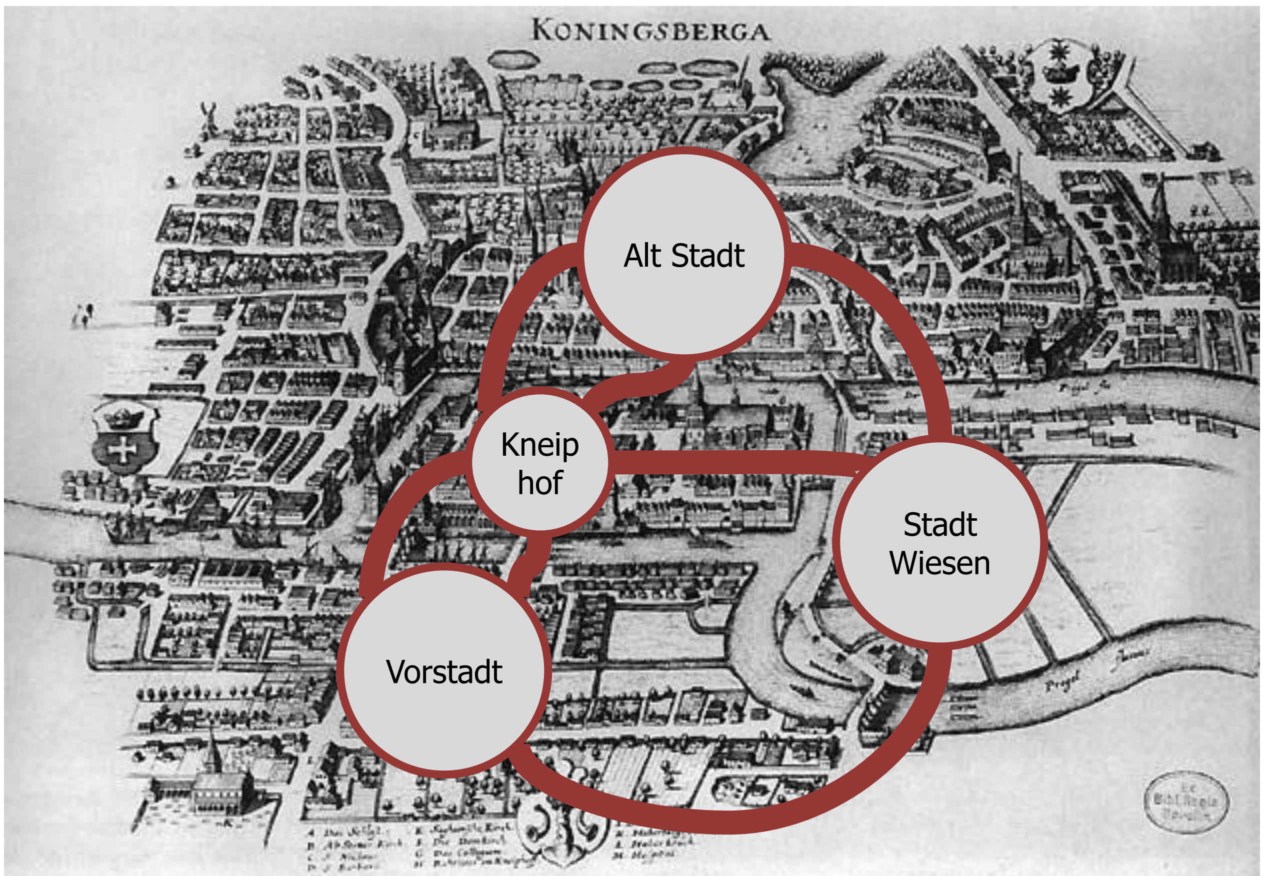
Overview

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?

Graph 101

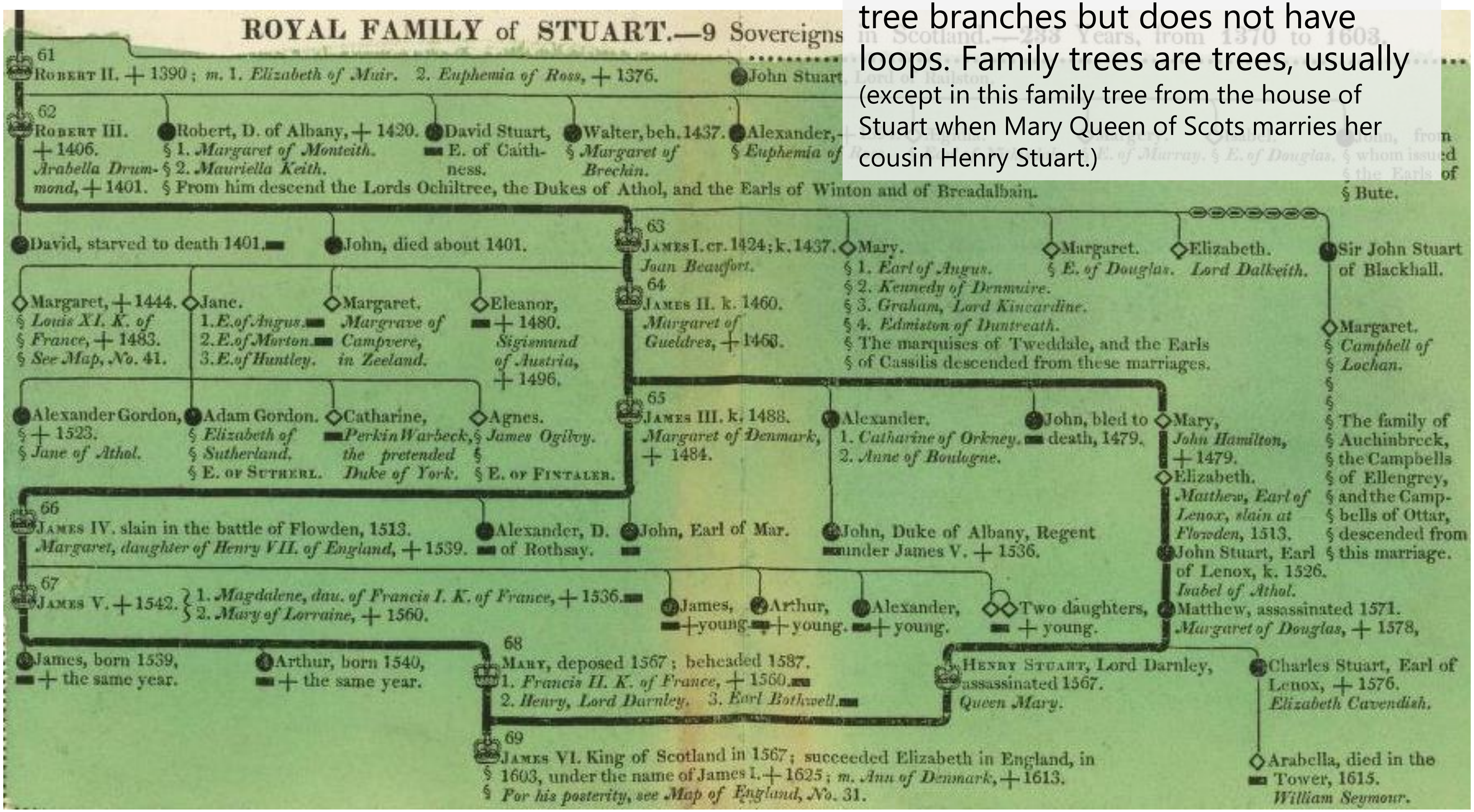
Seven Bridges of Königsberg problem (1735 – Leonhard Euler)

- Find a walk through the city that crosses each bridge once and only once.



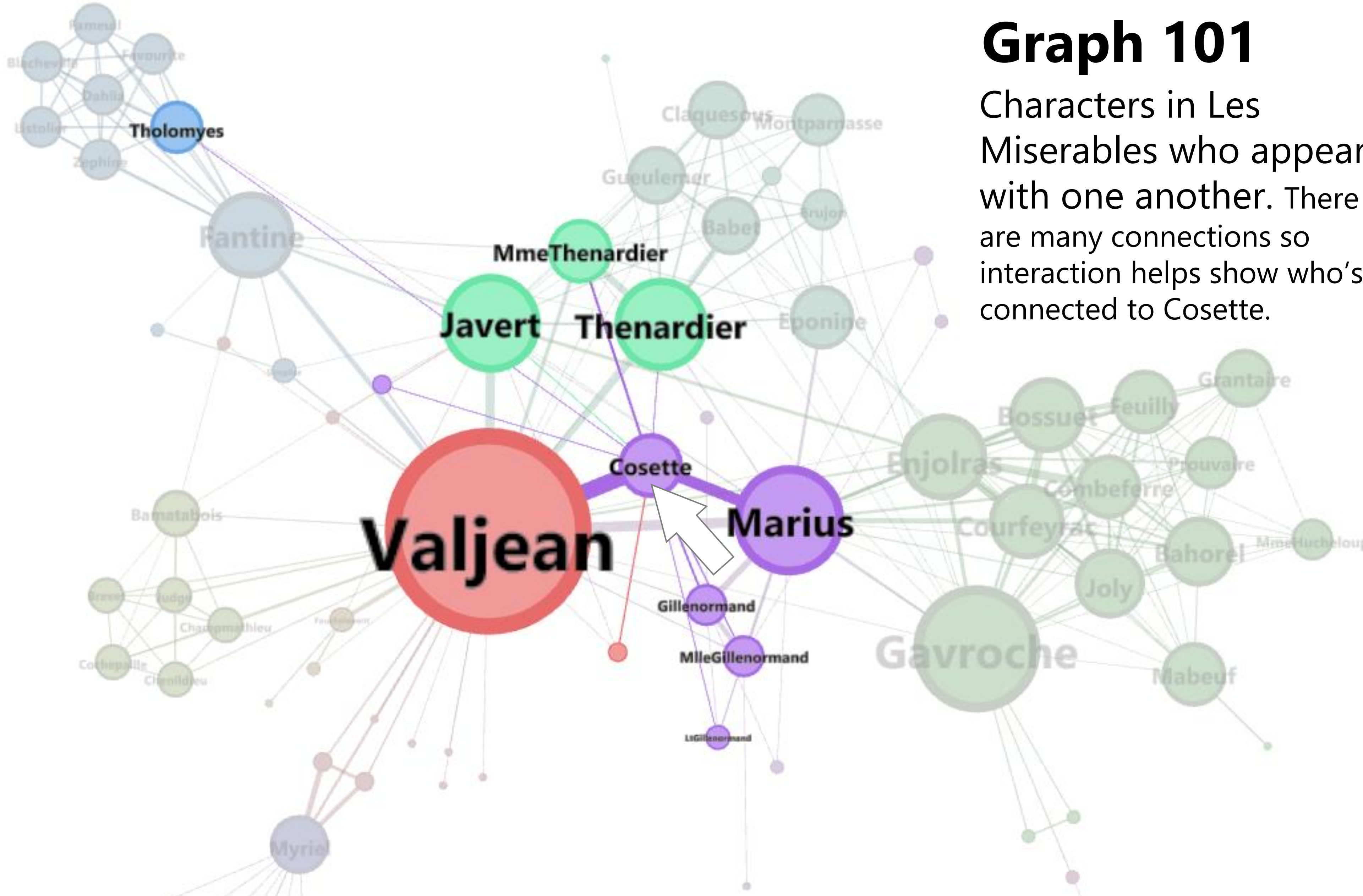
Graph 101

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.)



A Complete Genealogical, Historical, Chronological, And Geographical Atlas; Being A General Guide To History, Both Ancient And Modern According To The Plan Of Le Sage, Greatly Improved. The Whole Forming A Complete System Of History And Geography. By M. Lavoisne. Published By M. Carey And Son. 1820. Image from the David Rumsey Map Collection.
http://www.davidrumsey.com/luna/servlet/detail/RUMSEY~8~1~35452~1200370:Genealogical,-Historical,-and-Chron?sort=Pub_List_No_InitialSort%2CPub_Date%2CPub_List_No%2CSeries_No&qvq=q:genealogy;sort:Pub_List_No_InitialSort%2CPub_Date%2CPub_List_No%2CSeries_No;lc:RUMSEY~8~1&mi=12&trs=30

Graph 101



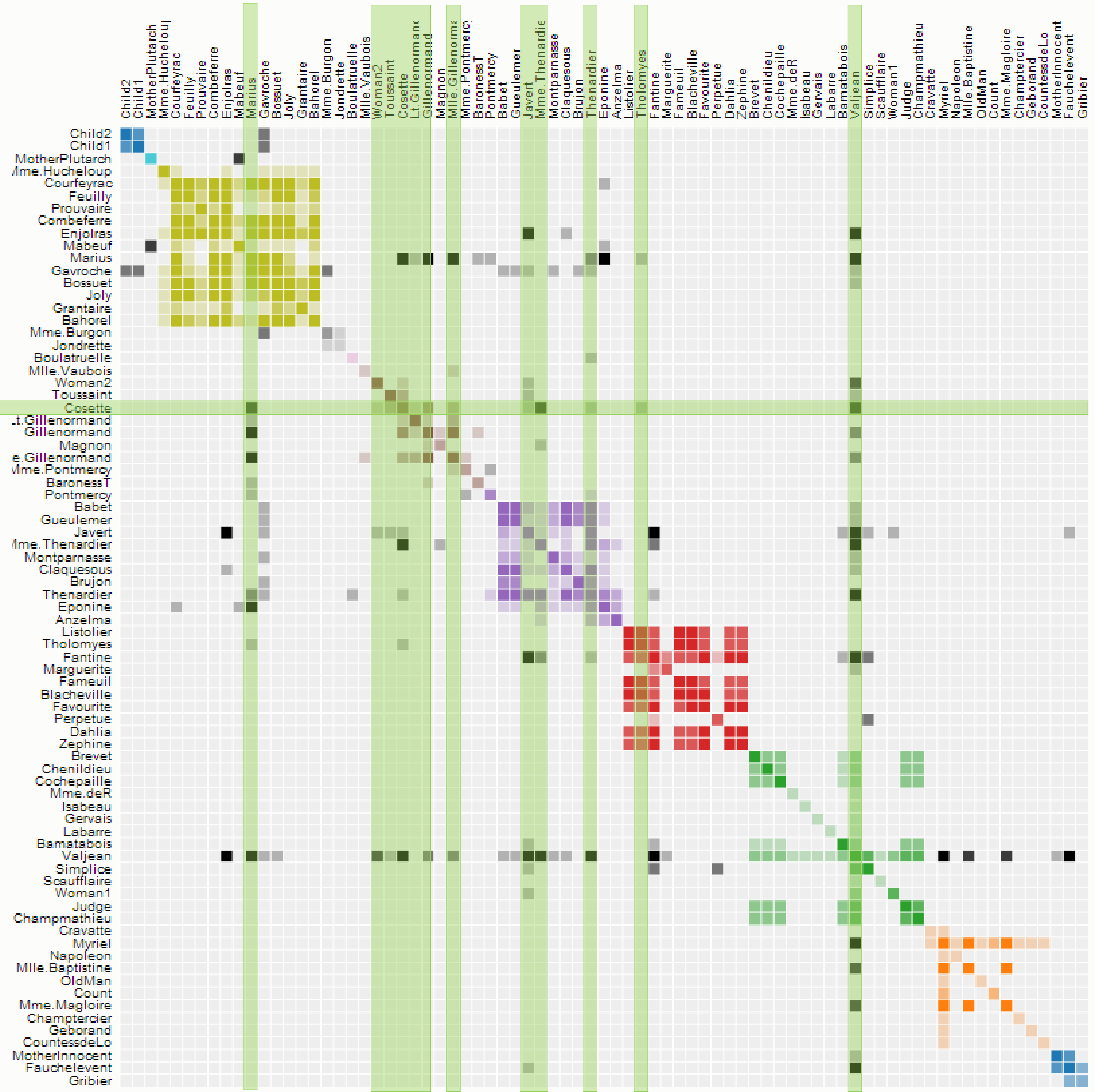
Characters in Les Miserables who appear with one another. There are many connections so interaction helps show who's connected to Cosette.

from Gephi: open-source point-and click graph visualization. Download it after this talk! www.gephi.org

Graph 101

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.

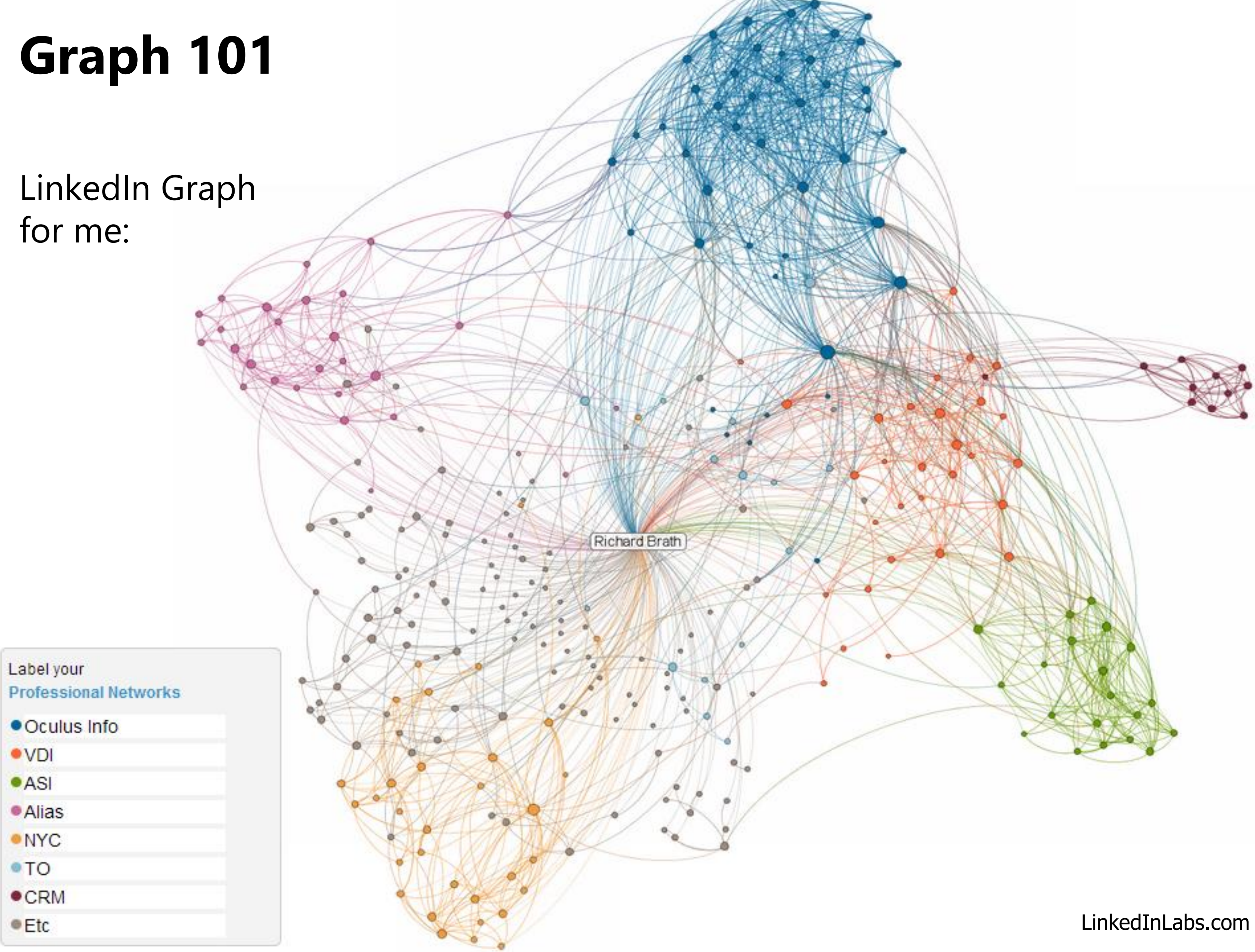


The row for Cosette has been highlighted with the corresponding connections to other nodes highlighted.

From: http://bost.ocks.org/mike/miserables/

Graph 101

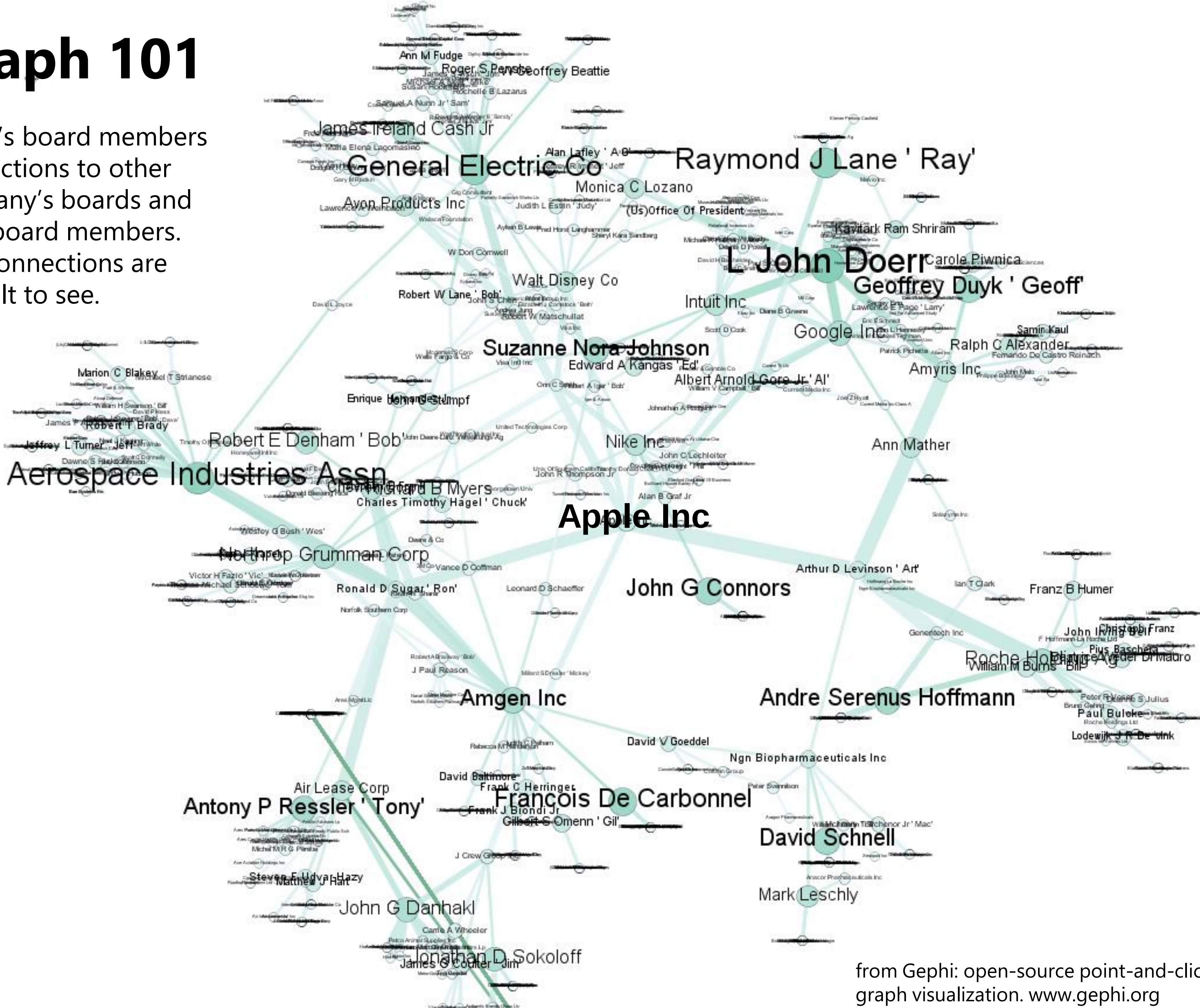
LinkedIn Graph
for me:



LinkedInLabs.com

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

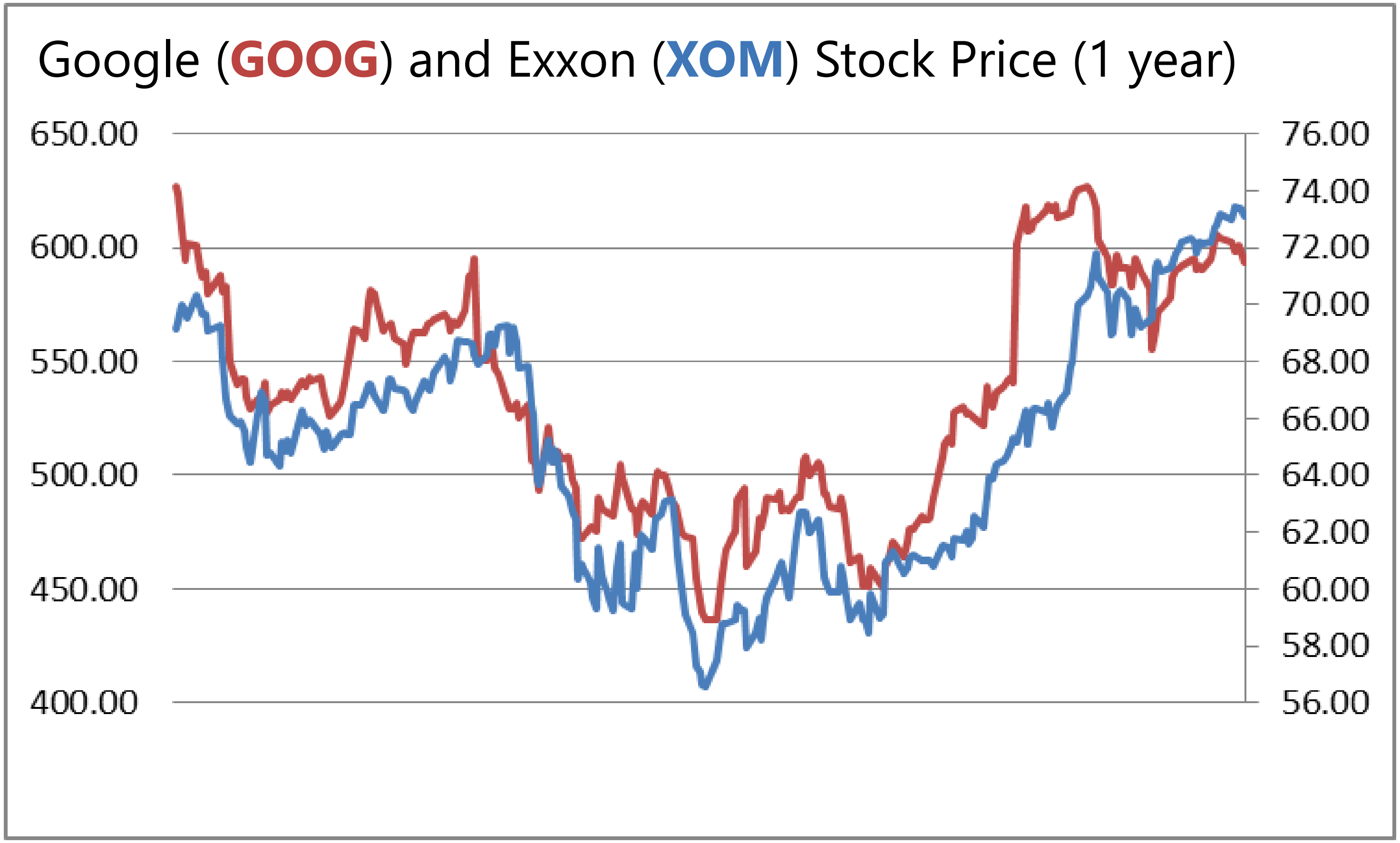
Who’s close?
Who’s opposite?

→ correlation graph

Visualizing a Correlation Graph

Step 1. Timeseries Correlation

Any two timeseries of data may have similar movement:

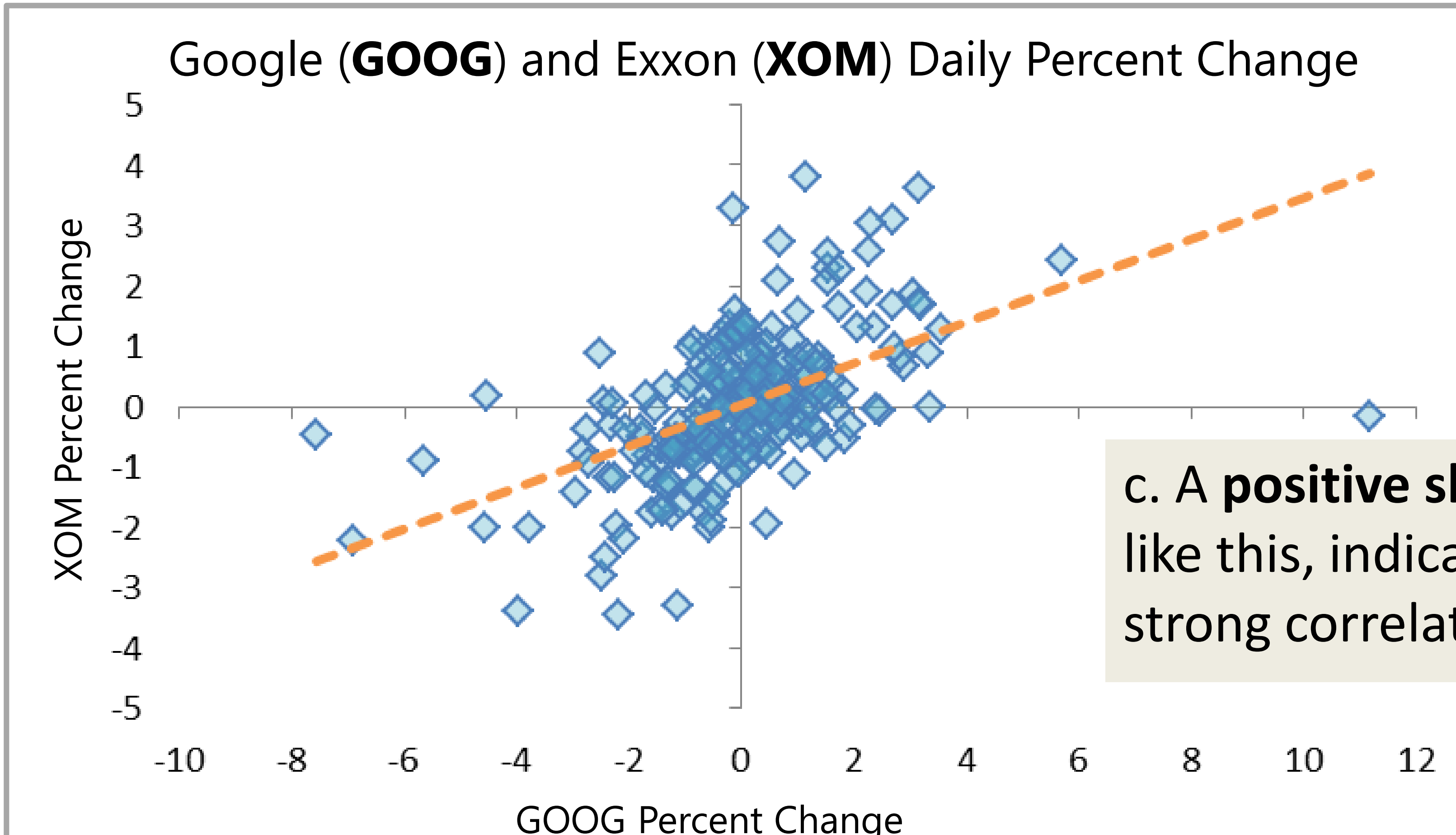


Visualizing a Correlation Graph

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.

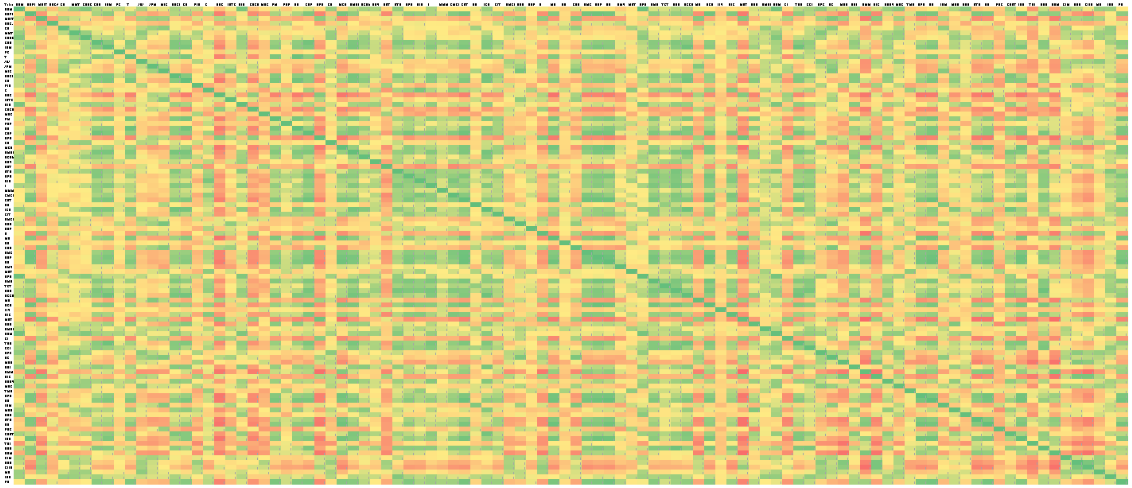
Ticker	XOM	GOOG	CVX	IBM	PG	AAPL	MSFT
XOM		0.892	0.700	0.639	0.677	0.268	0.640
GOOG	0.892		0.667	0.688	0.672	0.301	0.512
CVX	0.700	0.667		0.867	0.436	0.751	0.117
IBM	0.639	0.688	0.867		0.516	0.843	-0.075
PG	0.677	0.672	0.436	0.516		0.283	0.476
AAPL	0.268	0.301	0.751	0.843	0.283		-0.405
MSFT	0.640	0.512	0.117	-0.075	0.476	-0.405	

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

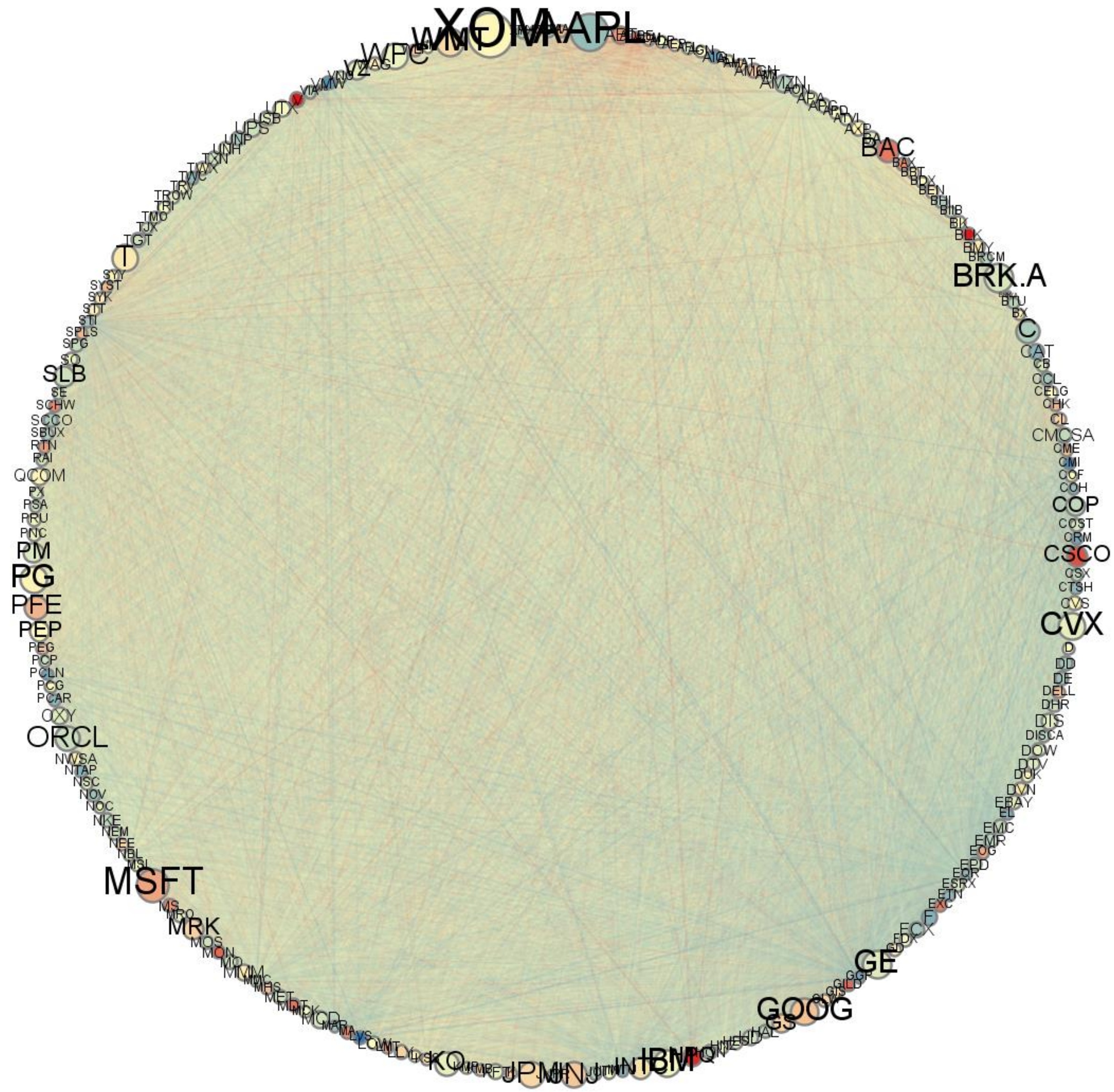
Correlations in a matrix. Difficult to scale:



With 200 items, there are now 40,000 cells (200x200)
> patterns such as clusters and outliers are not
easily discernable

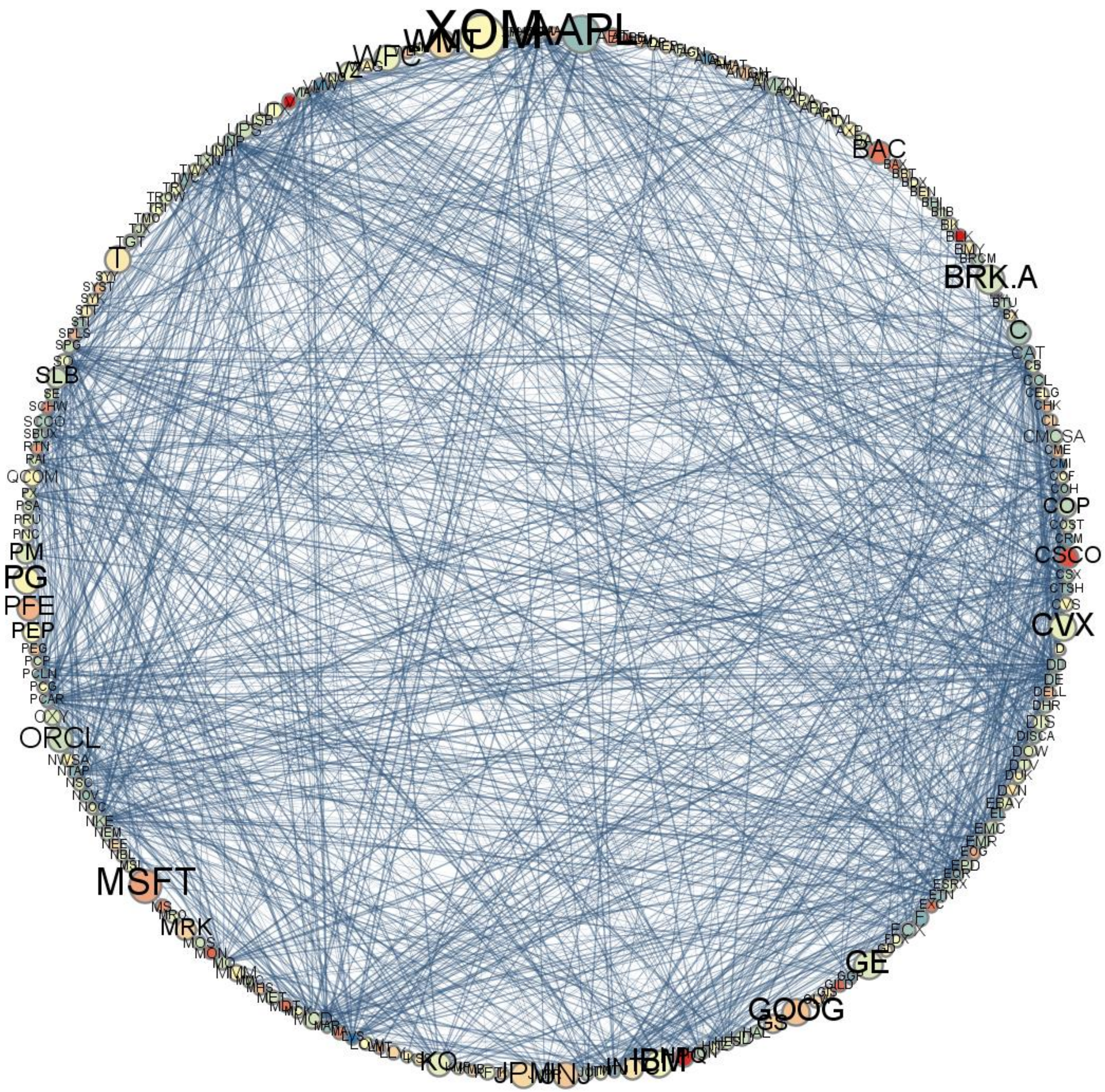
Instead: Draw it as a node + link graph:

- 200 stocks around perimeter
- Every correlation shown as a line
- Too many lines to gain any insight

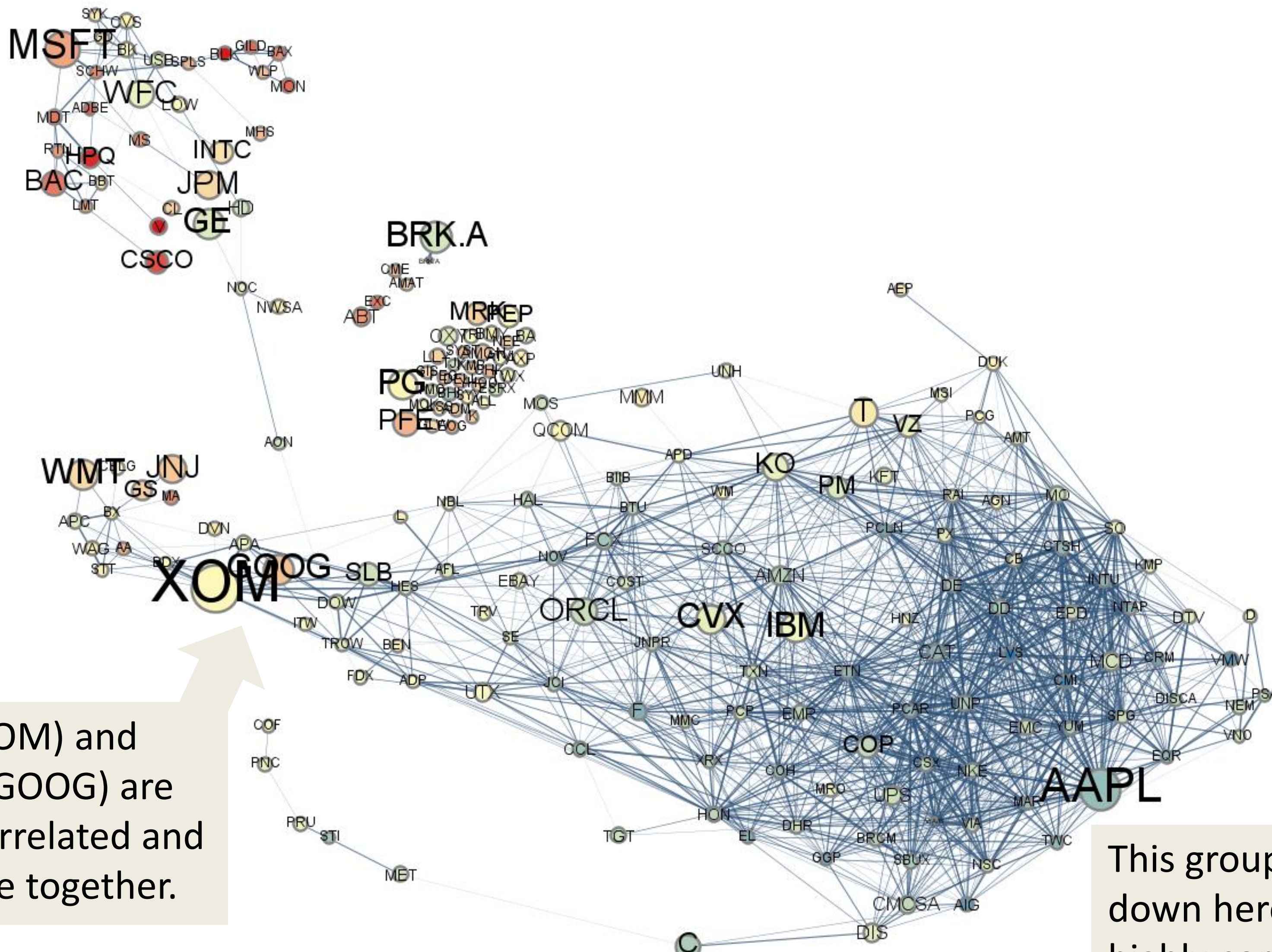


These are the strong connections…

- Remove most of the lines - only correlations between .85 - 1.0.
- Still too many lines to gain insight.



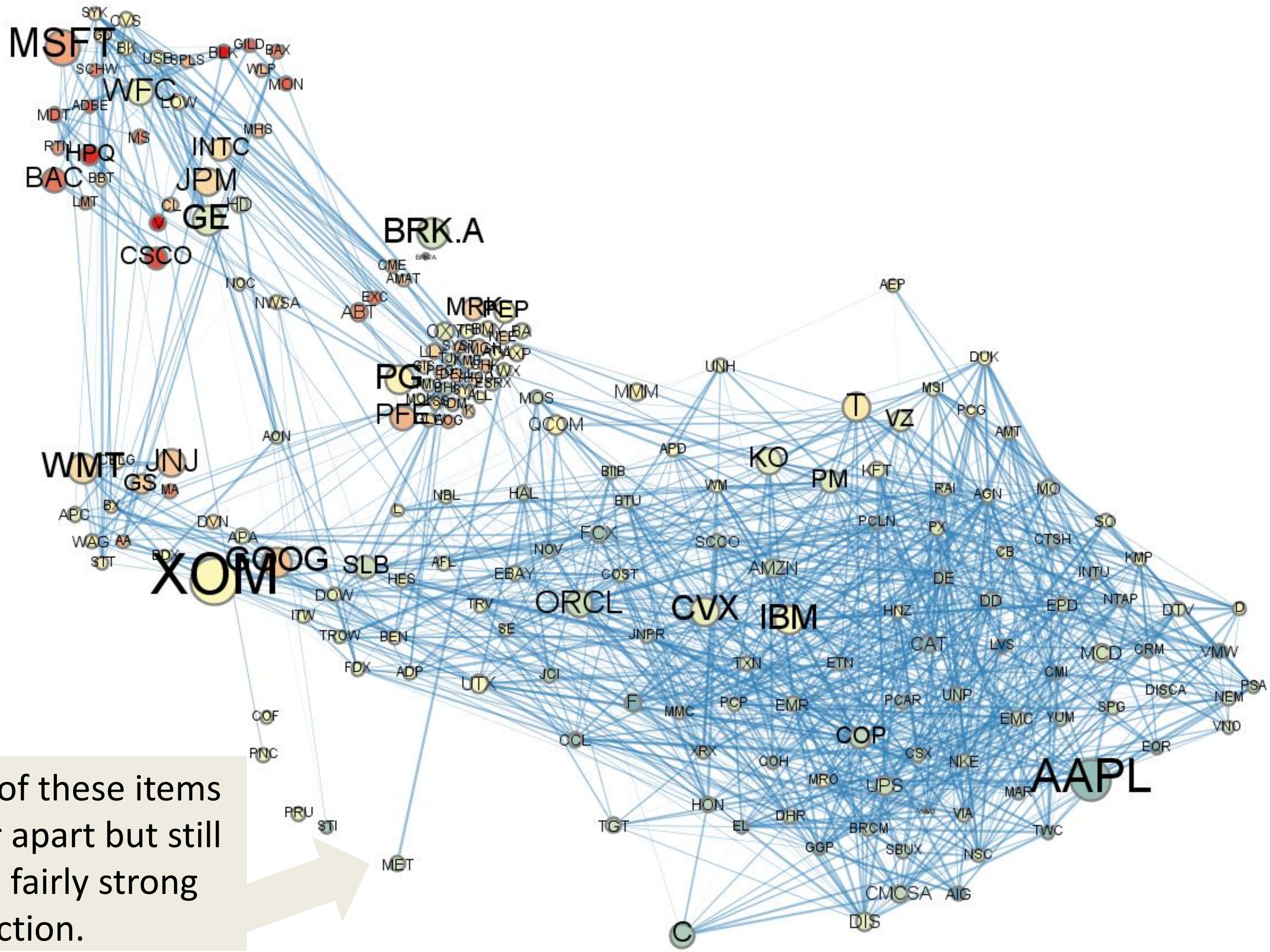
So pull together the strong connections, push everything else apart



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.

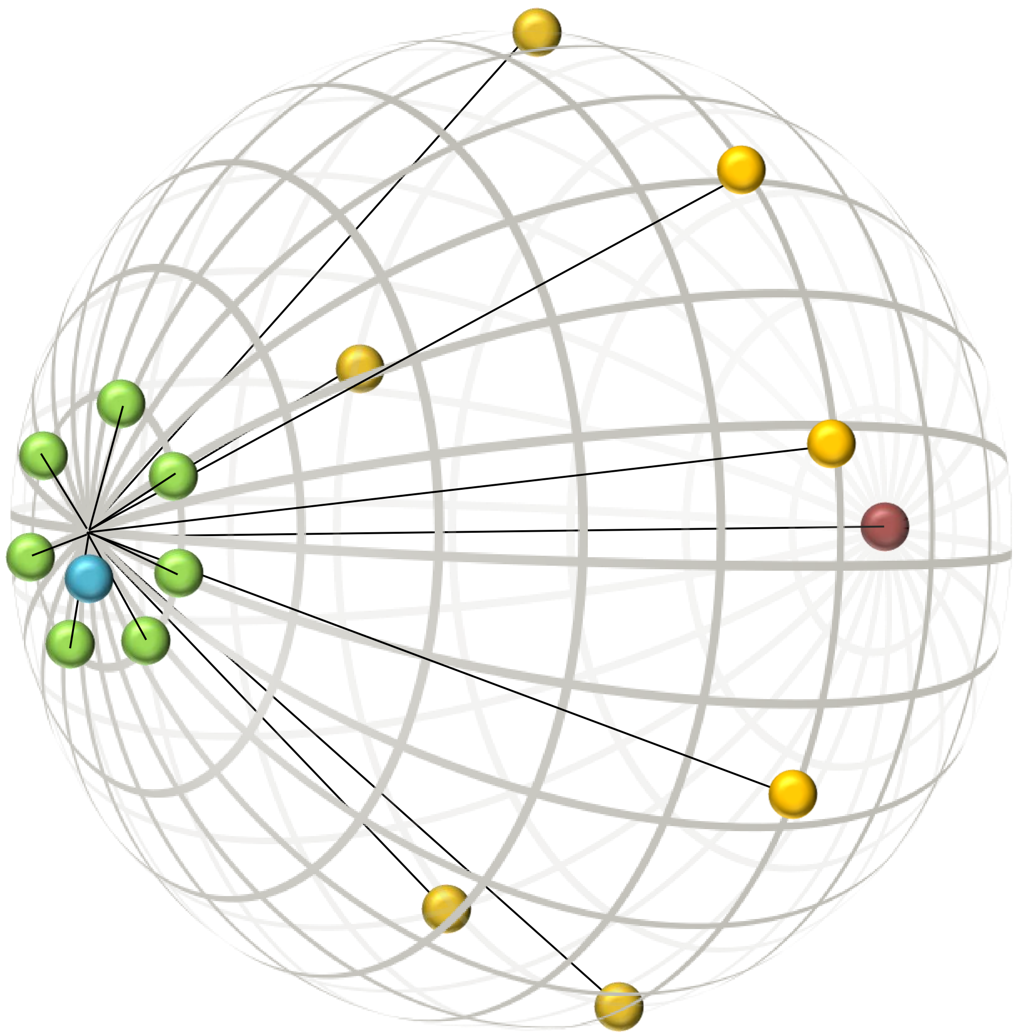
oculus

How about putting it on a 3D sphere?

Clustering on a sphere design premise –

For a given point:

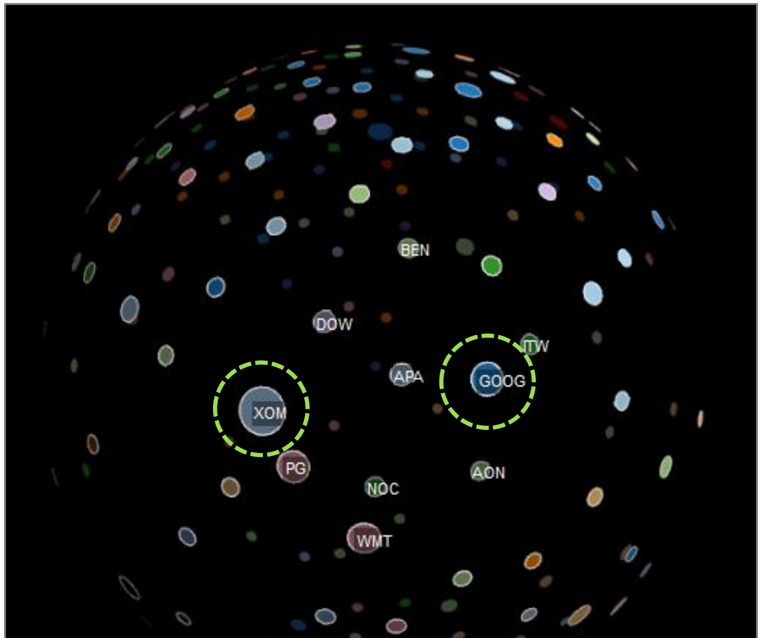
1. Strong correlations are close by.
2. Negative correlations are opposite.
3. Non-correlations are orthogonal.



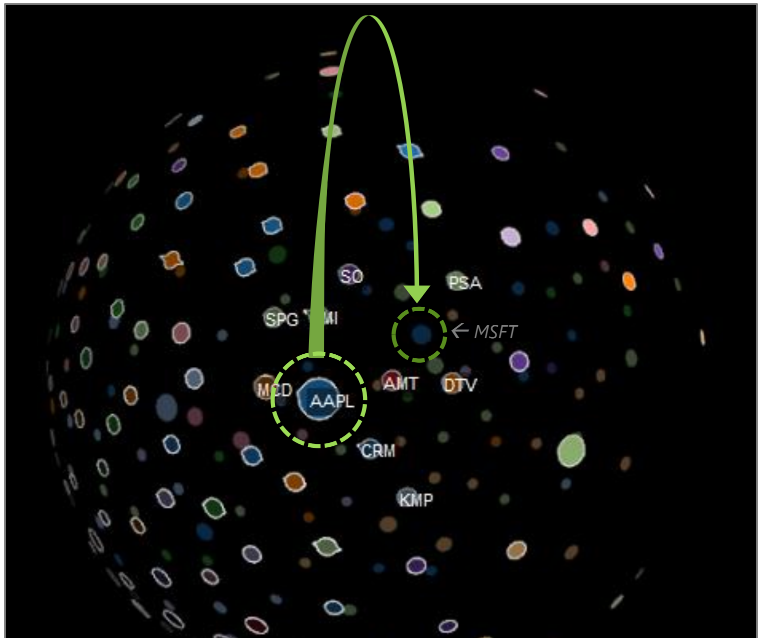
Sphere Correlation

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?

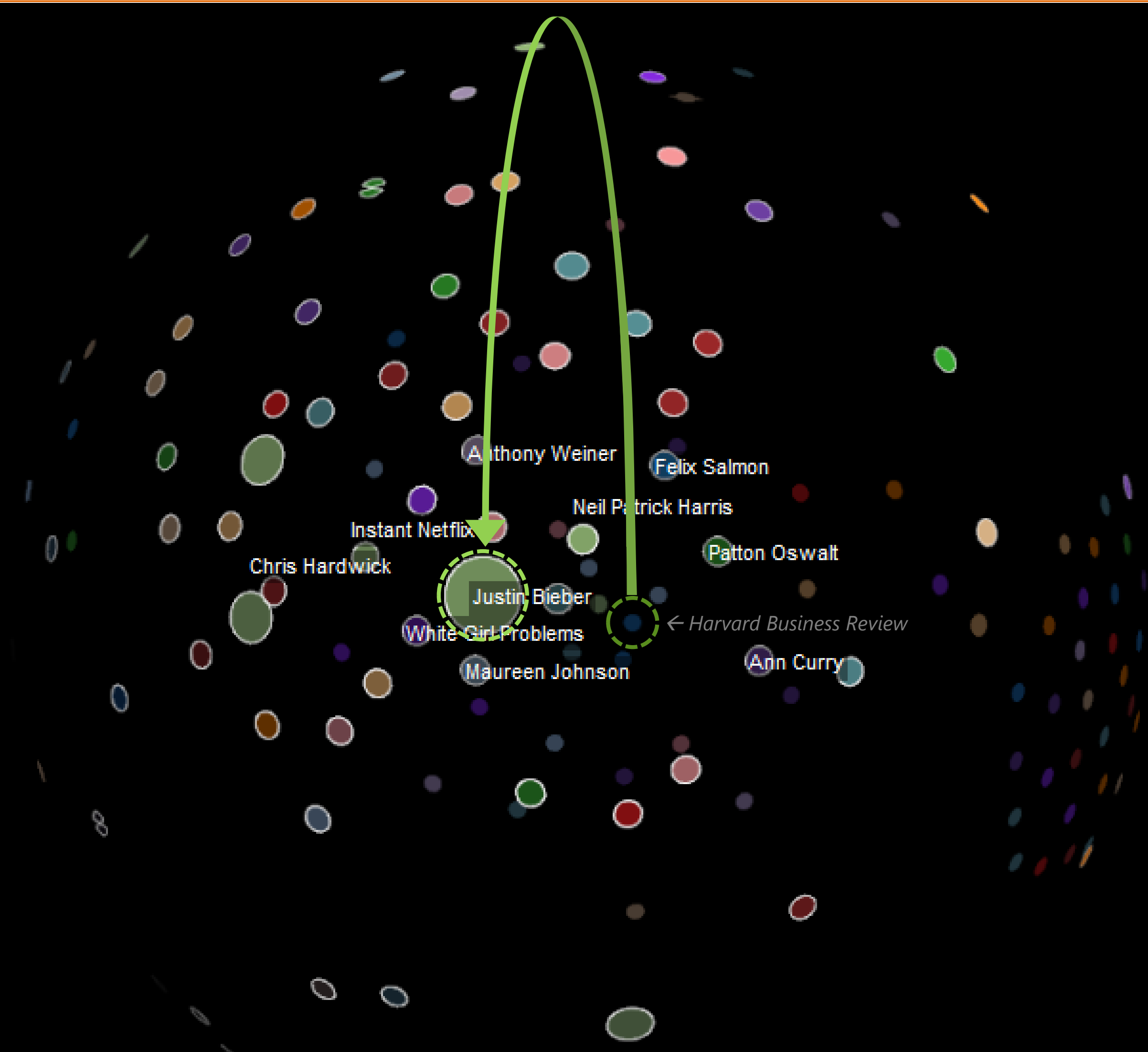


Exxon (XOM) and Google (GOOG) are near each other



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

Correlation Graph Summary

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

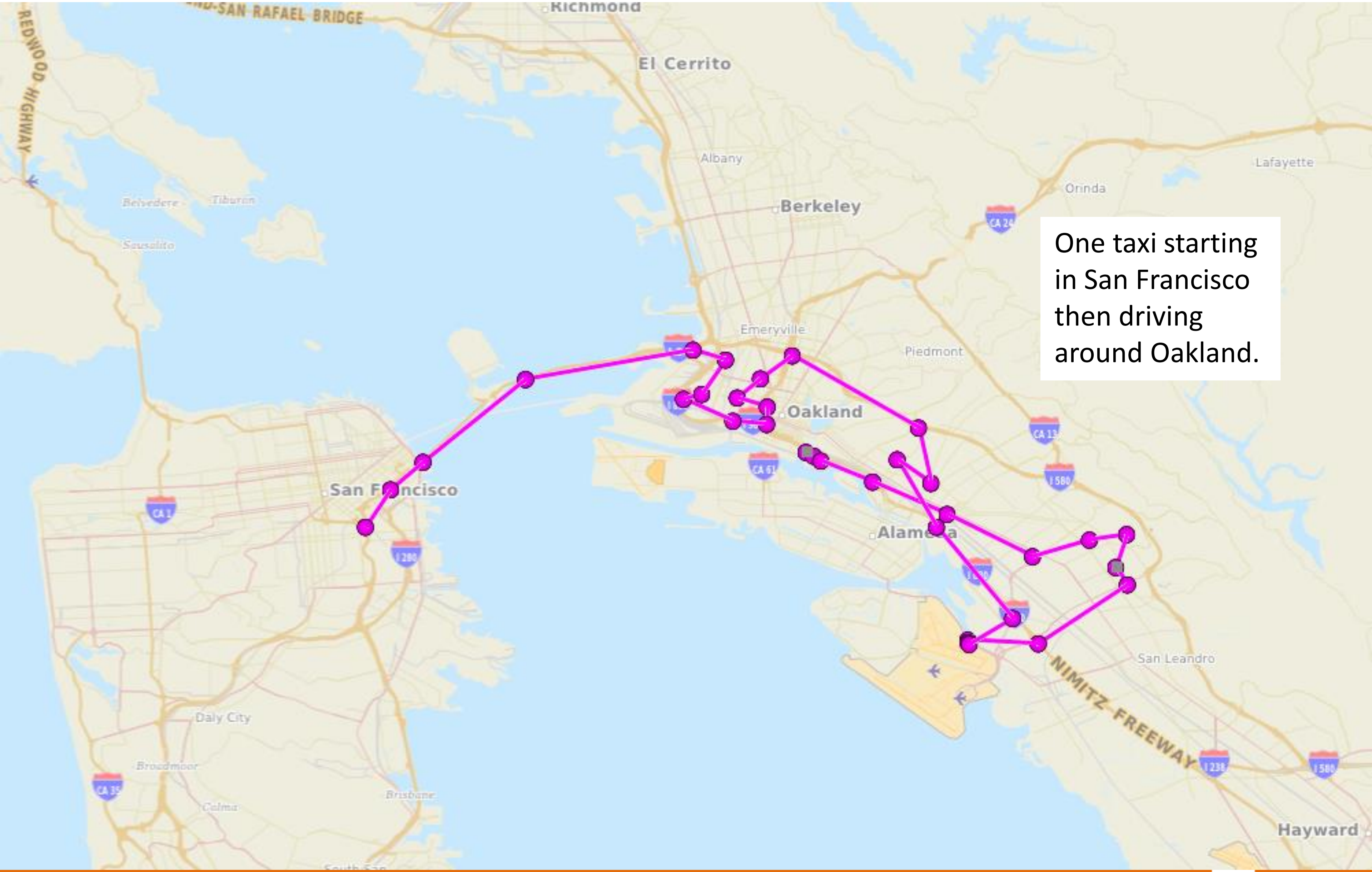
2. Use a layout that is both:
   + effective (in this case proximity)
   + intuitive (sphere works well for "opposite")

Making Sense of Big Graphs #2

Did these people meet?

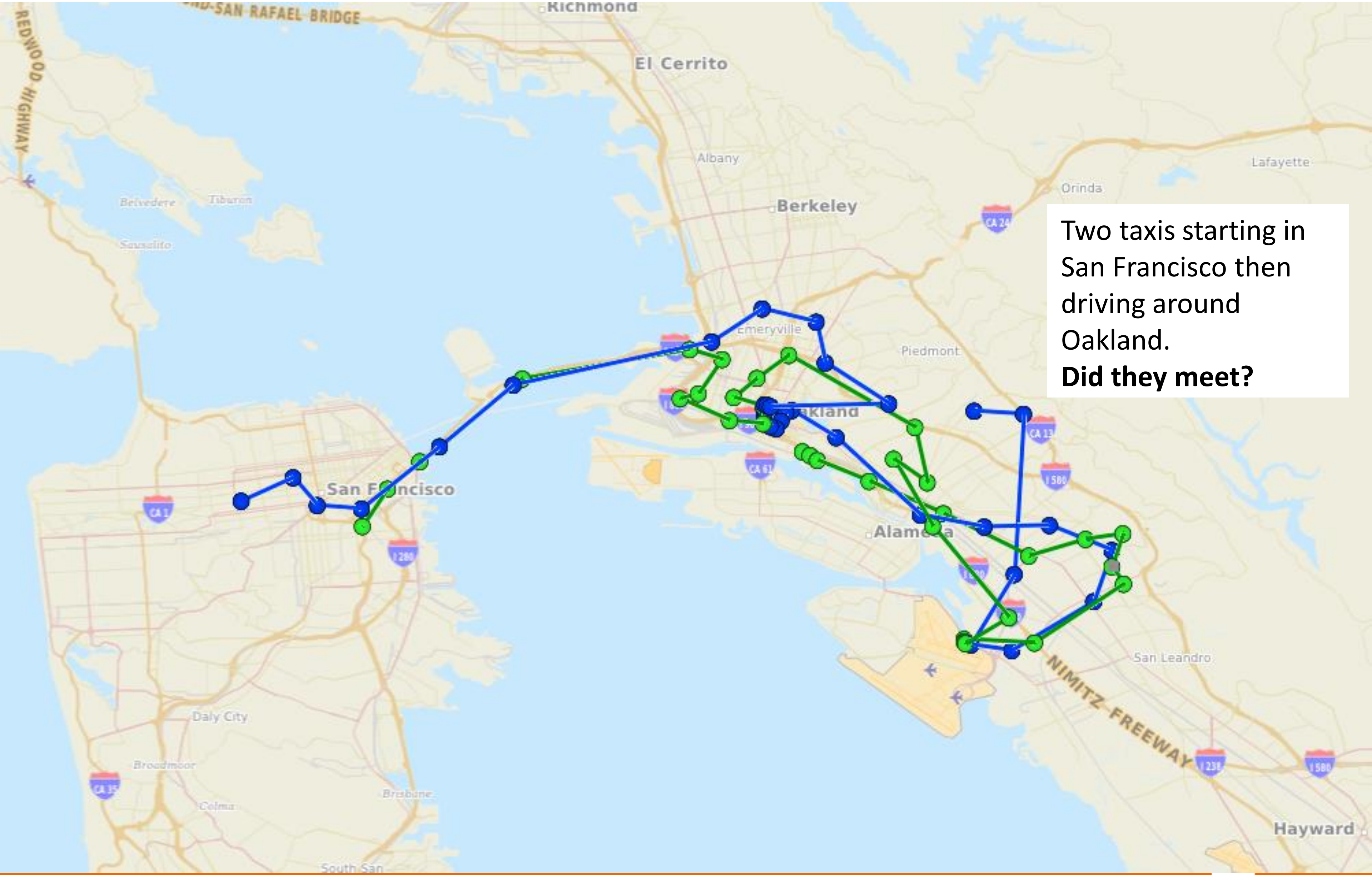
→ graphs in time and space

Graph on a map



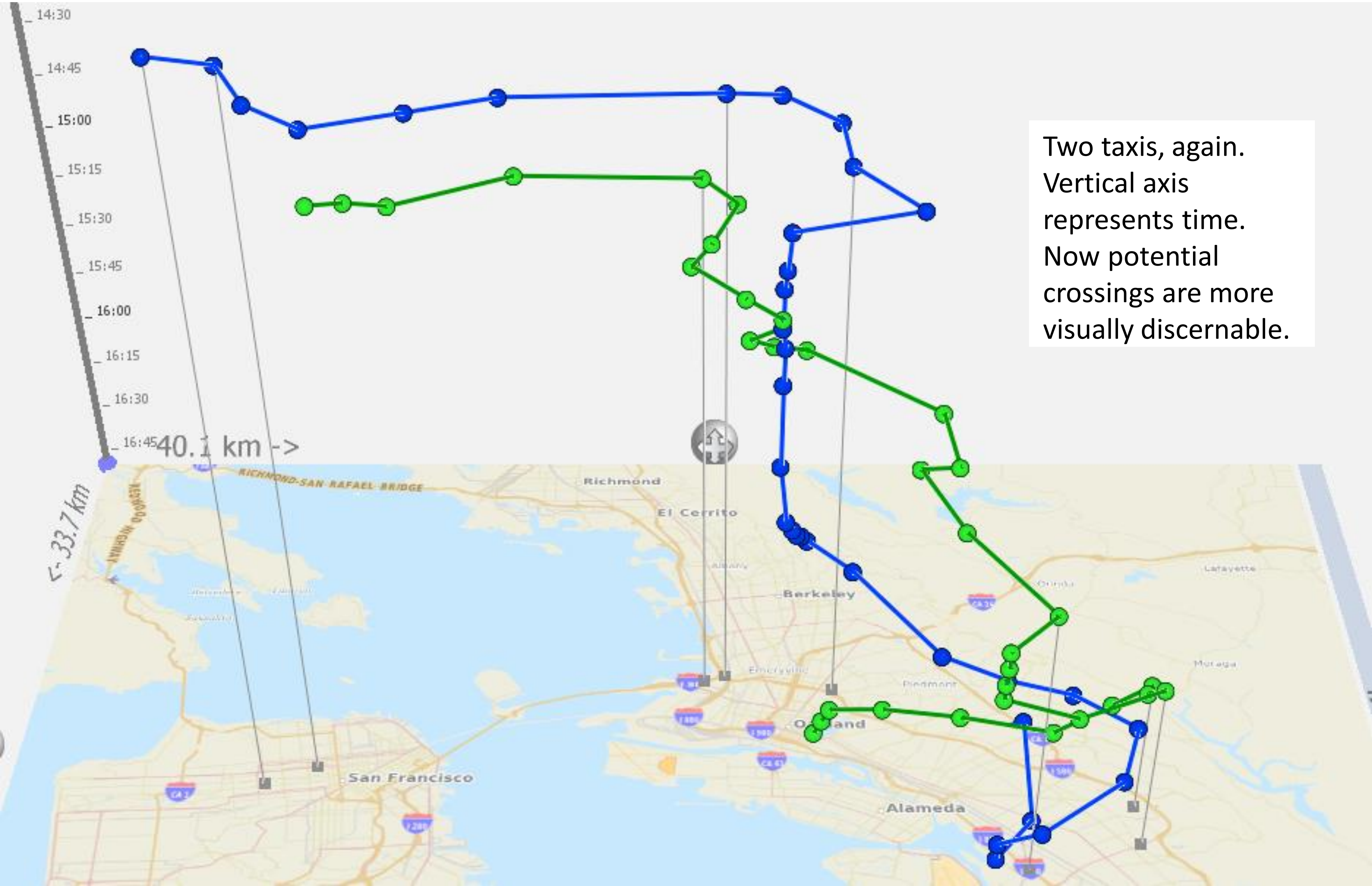
One taxi starting in San Francisco then driving around Oakland.

Graph on a map



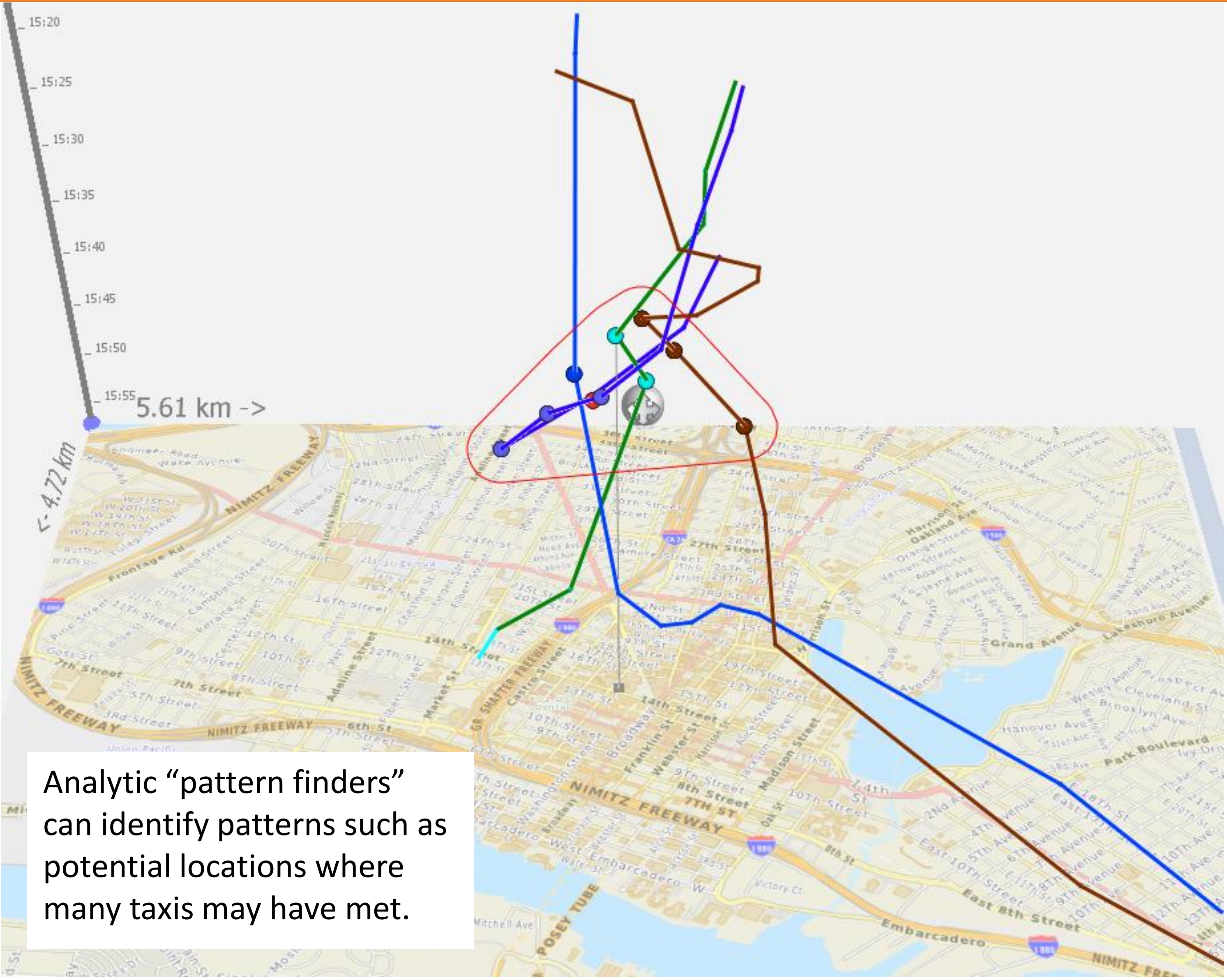
Two taxis starting in San Francisco then driving around Oakland.
Did they meet?

Graph on a map + time in third dimension



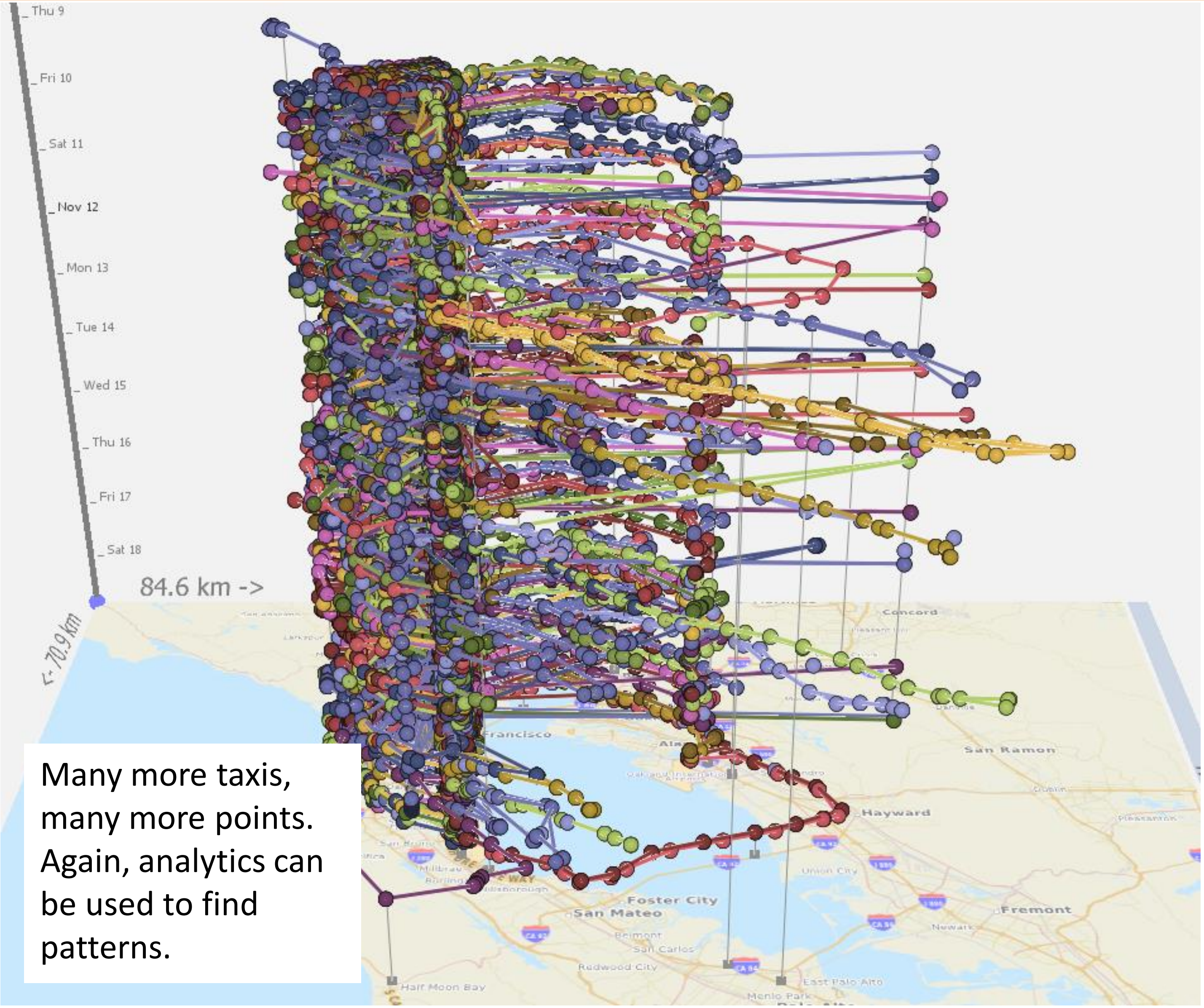
Two taxis, again. Vertical axis represents time. Now potential crossings are more visually discernable.

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.

oculus

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.

Charities by # of Donors	
9%	only 1 donor
63%	1 – 100 donors
23%	100 – 1000 donors
4%	1000 – 10000 donors
<1%	18 charities have 10000+

Charities by Total Donations	
29%	less than $10,000
38%	$10,000 - $100K
26%	$100K - $1M
6%	$1M - 10M
<1%	30 received $10M+

Donors	
67%	gave to 1 charity
32%	gave from 2-10
1%	gave to 11-100
< 1%	21 gave to 100+

Anonymized, public, open data.

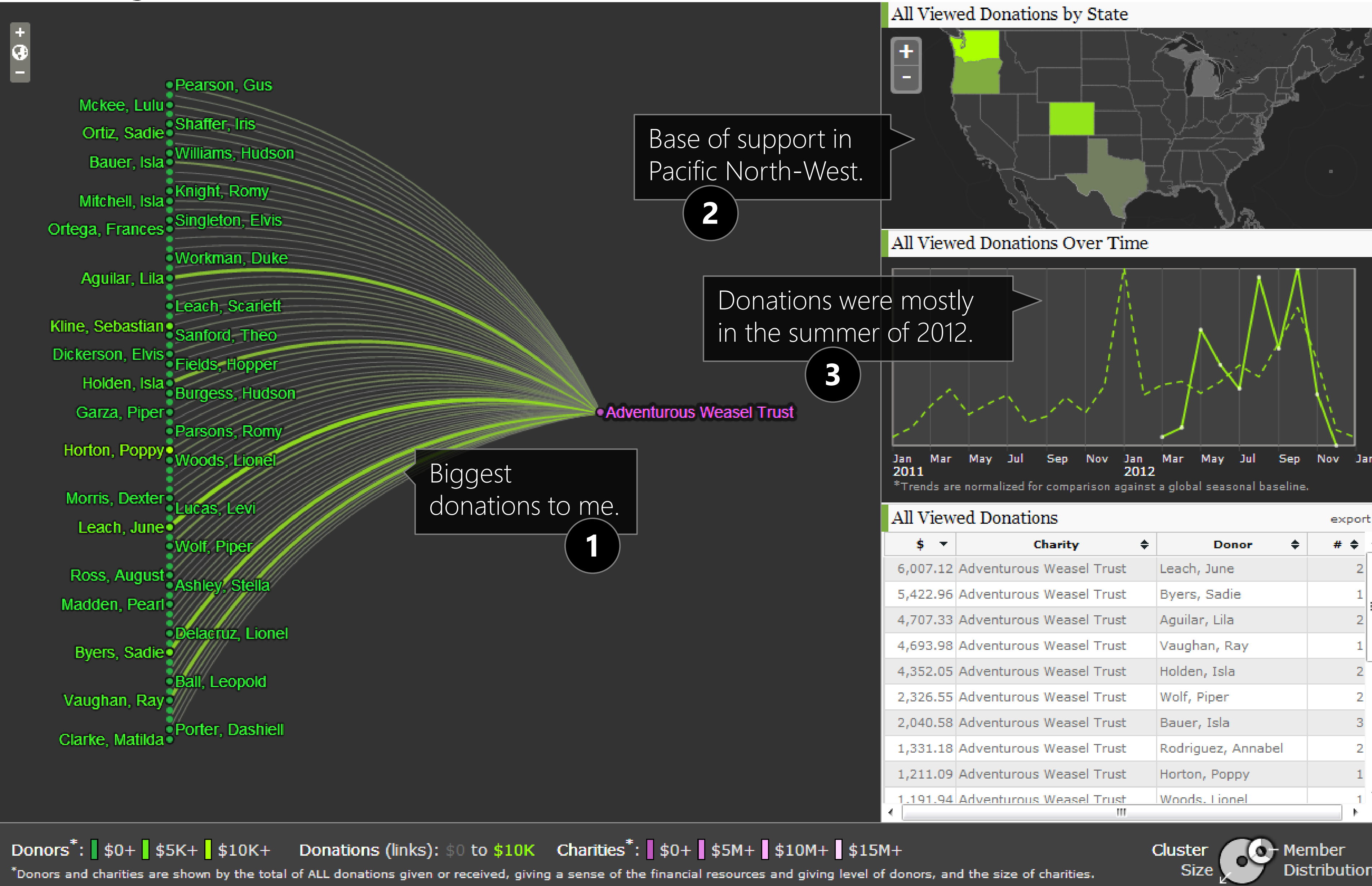
3 weeks to design and build an exploratory web-based tool.

For any charity:

Who are my donors?

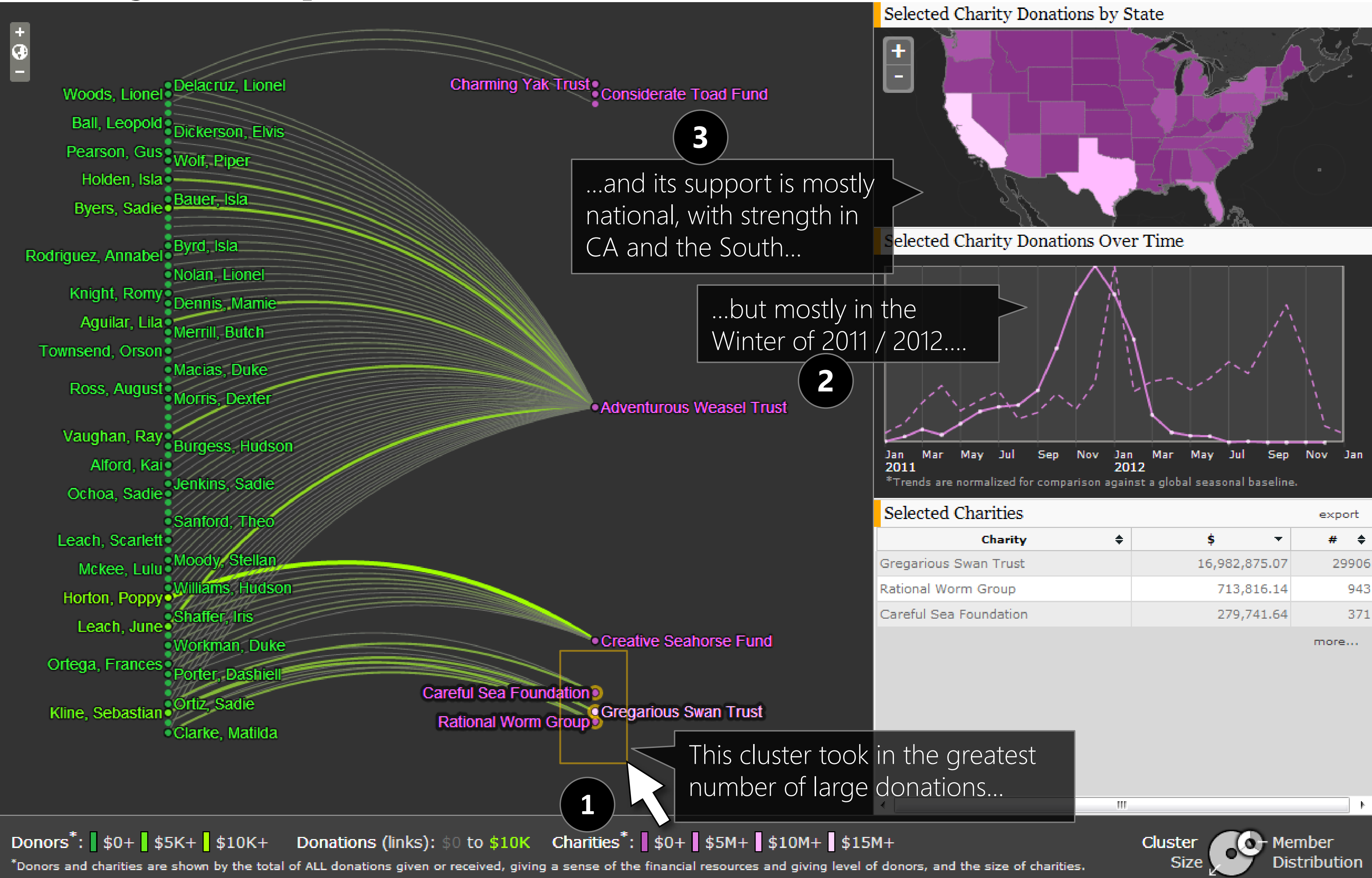
Who are my *prospective* donors?

My Donors



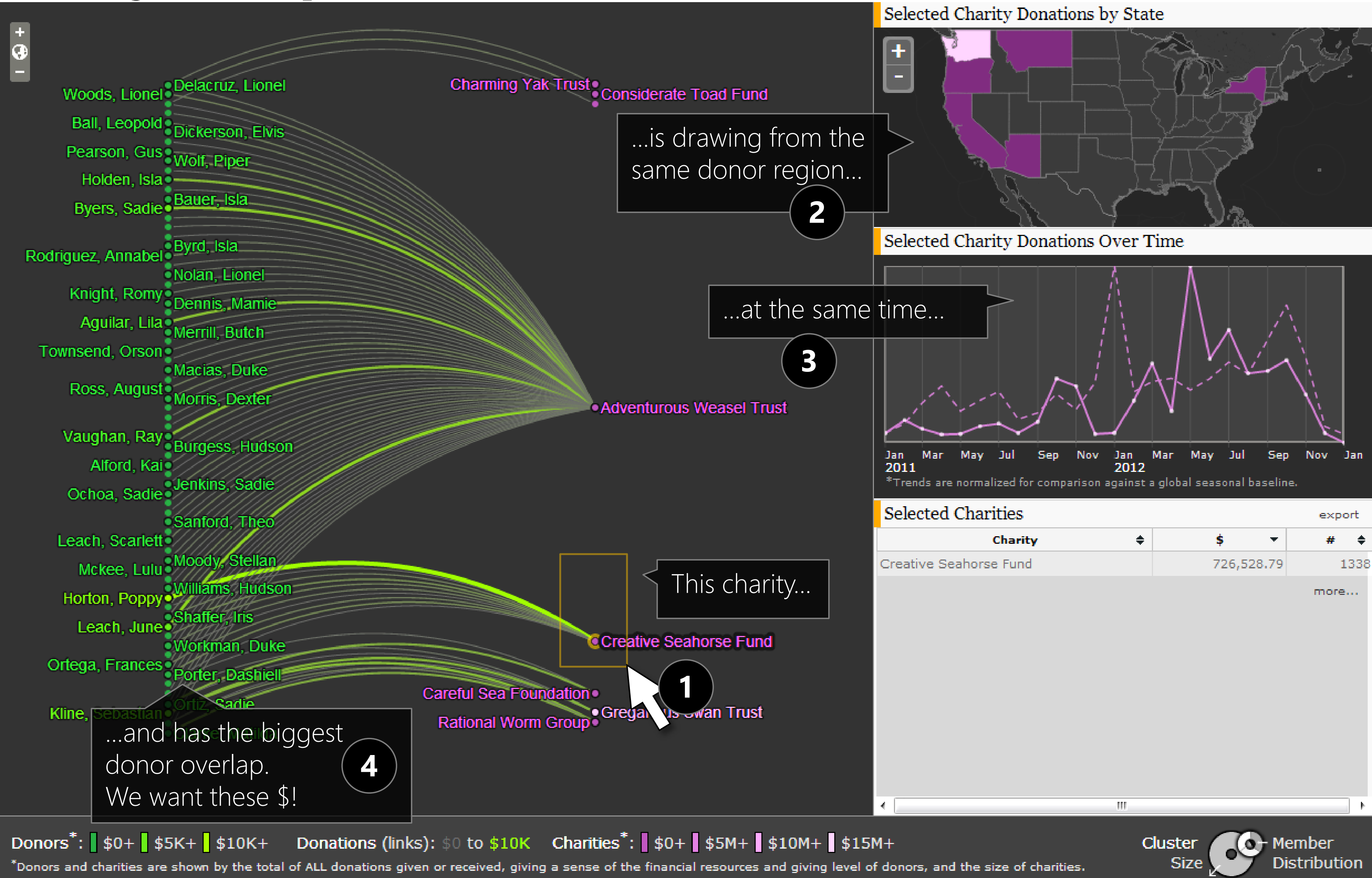
$	Charity	Donor	#
6,007.12	Adventurous Weasel Trust	Leach, June	2
5,422.96	Adventurous Weasel Trust	Byers, Sadie	1
4,707.33	Adventurous Weasel Trust	Aguilar, Lila	2
4,693.98	Adventurous Weasel Trust	Vaughan, Ray	1
4,352.05	Adventurous Weasel Trust	Holden, Isla	2
2,326.55	Adventurous Weasel Trust	Wolf, Piper	2
2,040.58	Adventurous Weasel Trust	Bauer, Isla	3
1,331.18	Adventurous Weasel Trust	Rodriguez, Annabel	2
1,211.09	Adventurous Weasel Trust	Horton, Poppy	1
1,191.94	Adventurous Weasel Trust	Woods, Lionel	1

My Competition?

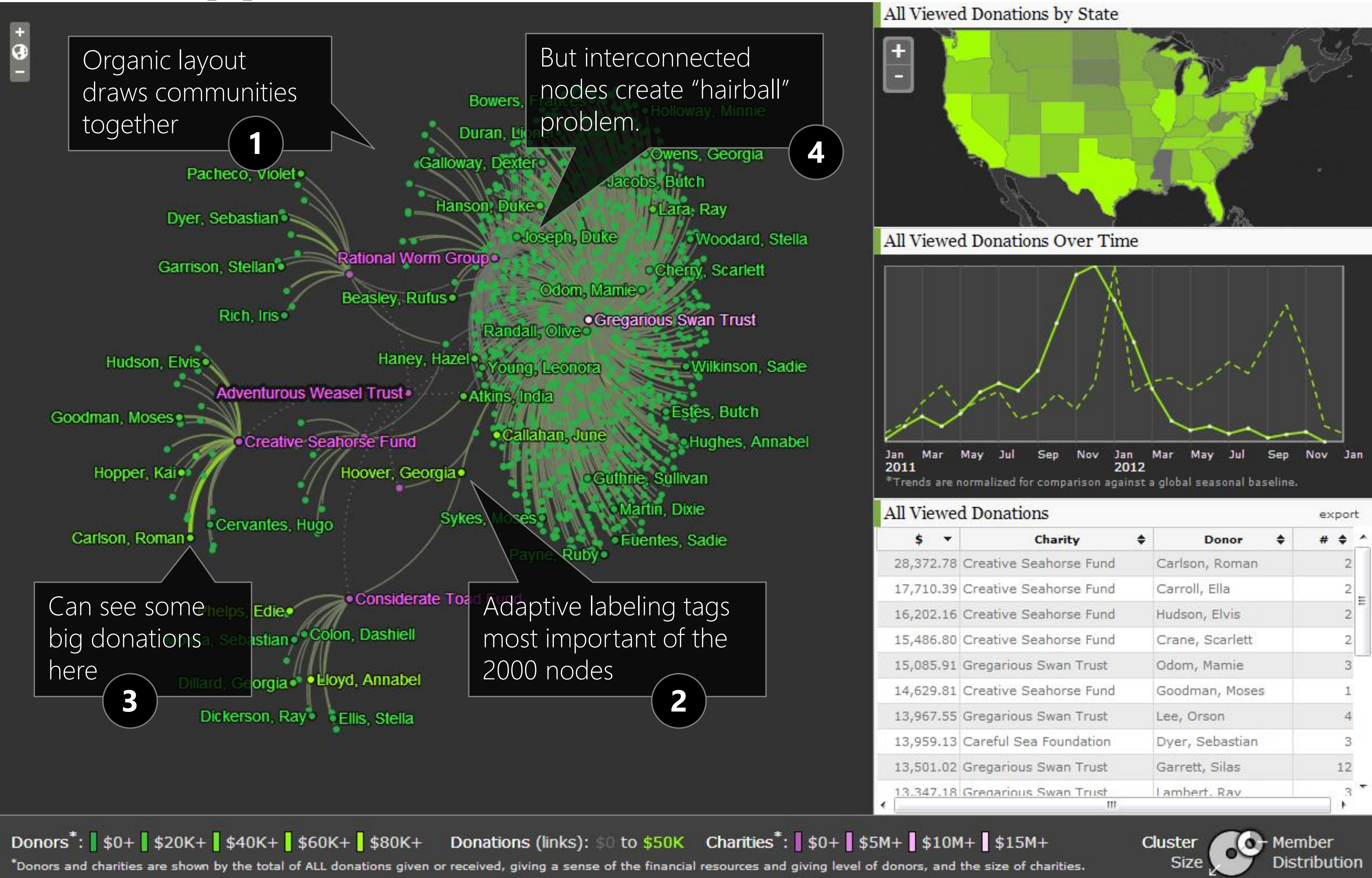


oculus

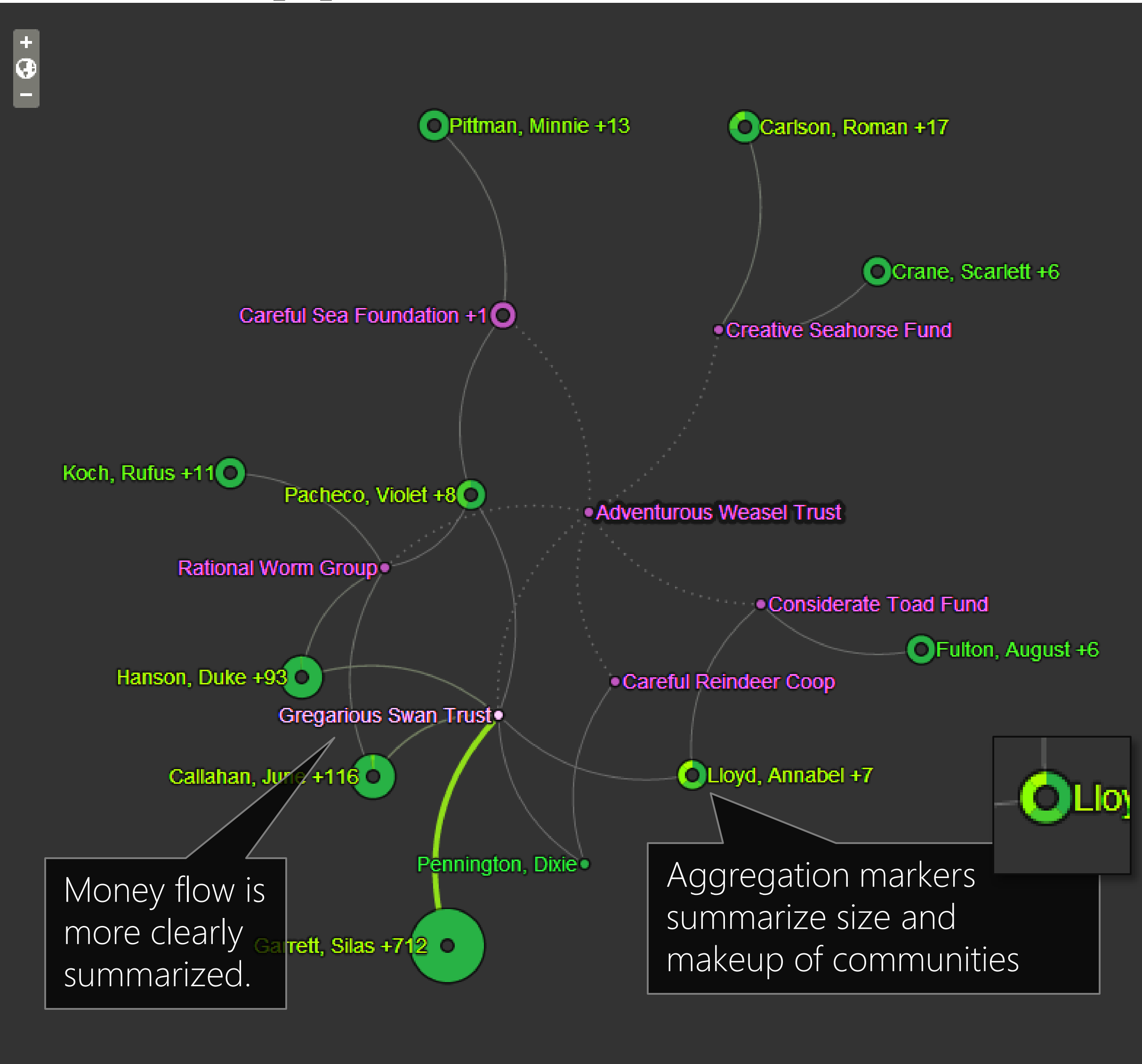
My Competition



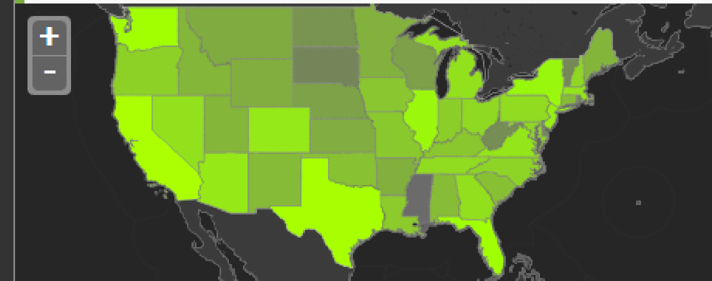
Untapped Donors?



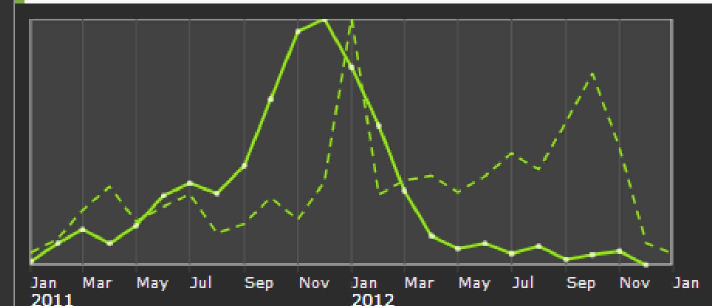
Untapped Donors: Community Aggregation



All Viewed Donations by State



All Viewed Donations Over Time



*Trends are normalized for comparison against a global seasonal baseline.

All Viewed Donations

export

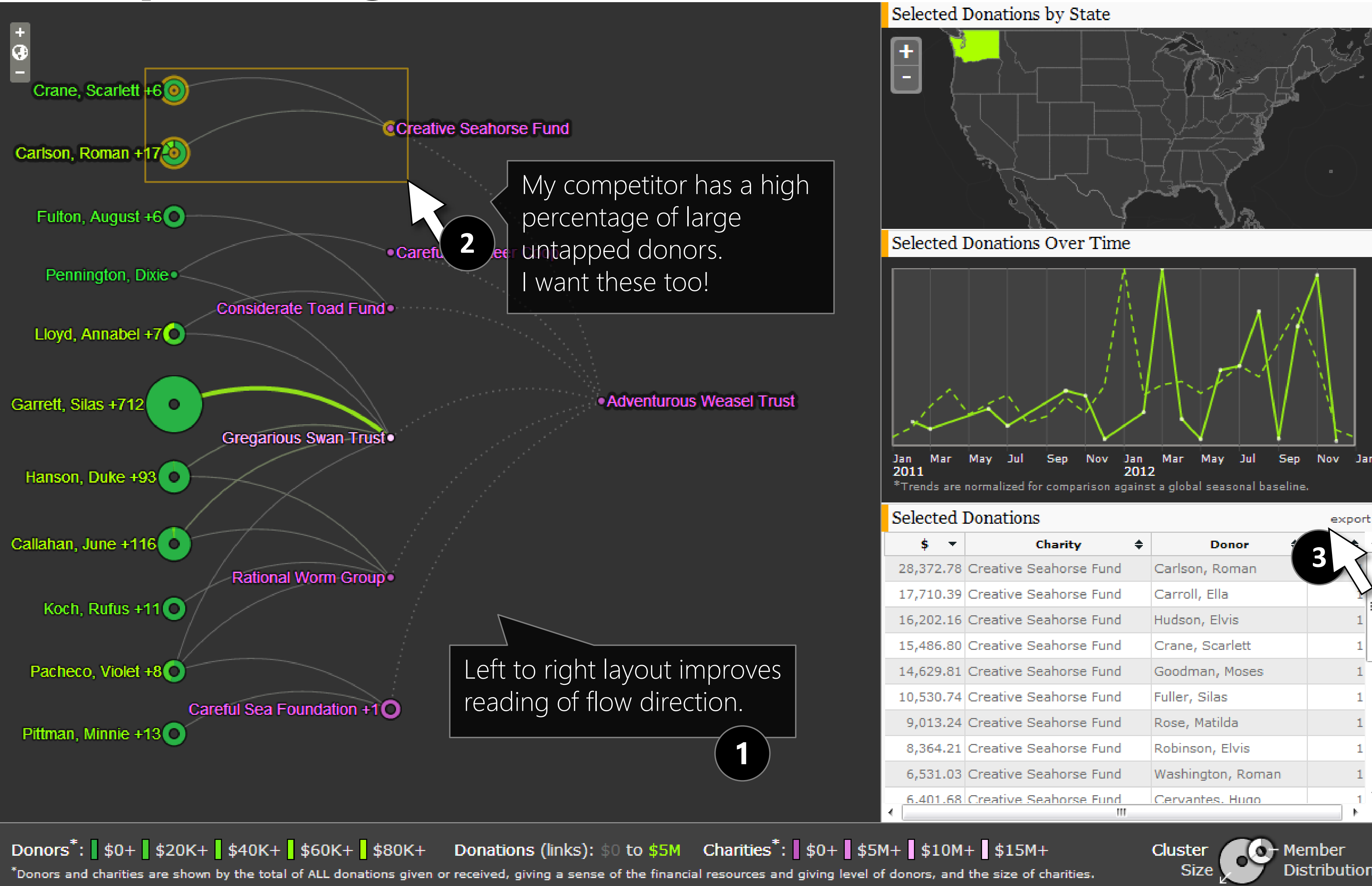
$	Charity	Donor	#
28,372.78	Creative Seahorse Fund	Carlson, Roman	2
17,710.39	Creative Seahorse Fund	Carroll, Ella	2
16,202.16	Creative Seahorse Fund	Hudson, Elvis	2
15,486.80	Creative Seahorse Fund	Crane, Scarlett	2
15,085.91	Gregarious Swan Trust	Odom, Mamie	3
14,629.81	Creative Seahorse Fund	Goodman, Moses	1
13,967.55	Gregarious Swan Trust	Lee, Orson	4
13,959.13	Careful Sea Foundation	Dyer, Sebastian	3
13,501.02	Gregarious Swan Trust	Garrett, Silas	12
13,347.18	Gregarious Swan Trust	Lambert, Ray	3

Donors*: $0+ $20K+ $40K+ $60K+ $80K+ Donations (links): $0 to $5M Charities*: $0+ $5M+ $10M+ $15M+ Cluster Size Member Distribution

*Donors and charities are shown by the total of ALL donations given or received, giving a sense of the financial resources and giving level of donors, and the size of charities.

oculus

Export Target Donors



Charity Network: Summary

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

- **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.

20,000 nodes with Louvain[3] clustering and aggregation markers

[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

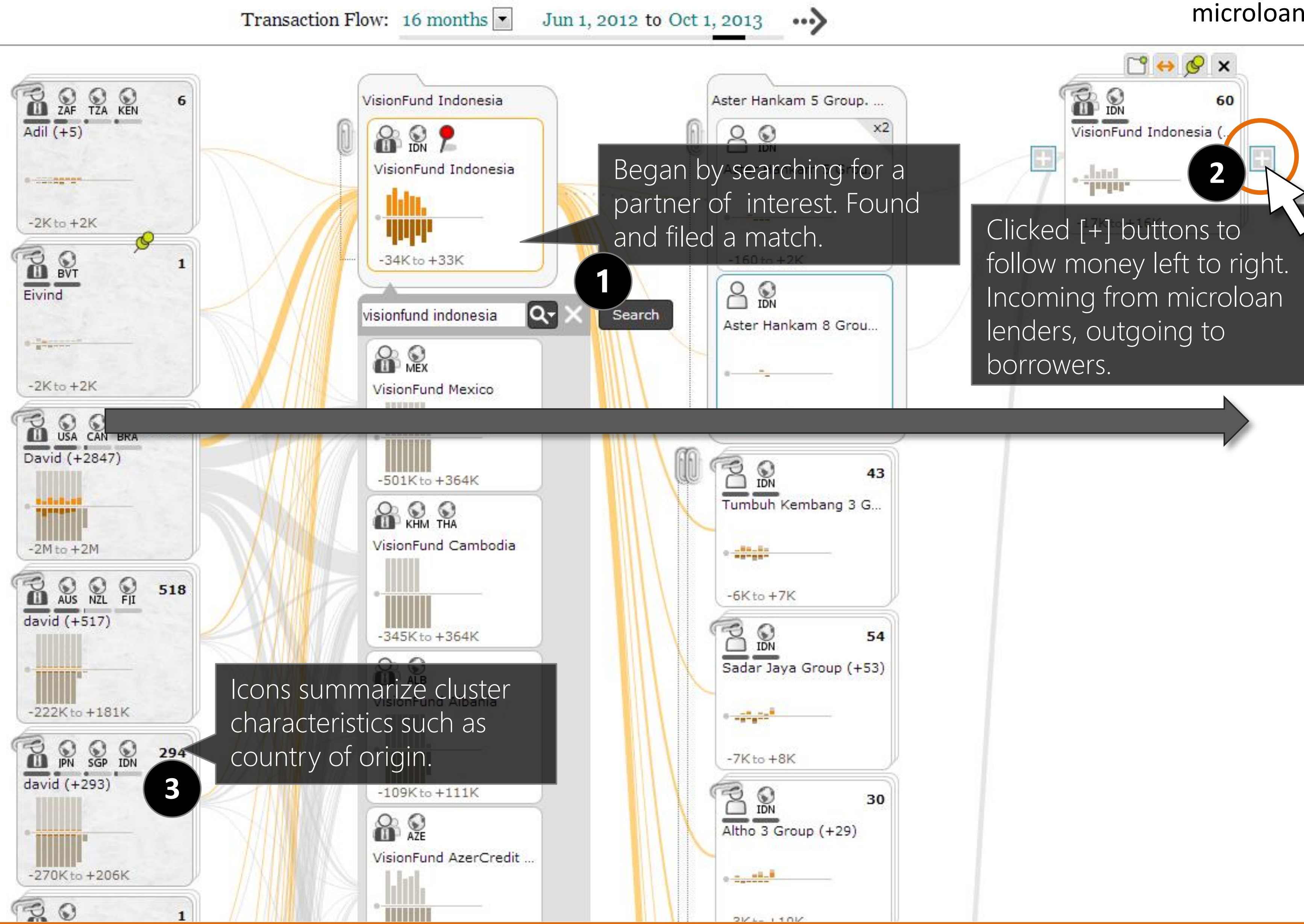
- **2,000,000** accounts. **4,000,000** transactions.

bitcoin

- **6,000,000** accounts. **16,000,000** transactions.

Example: Kiva Partner Transactions

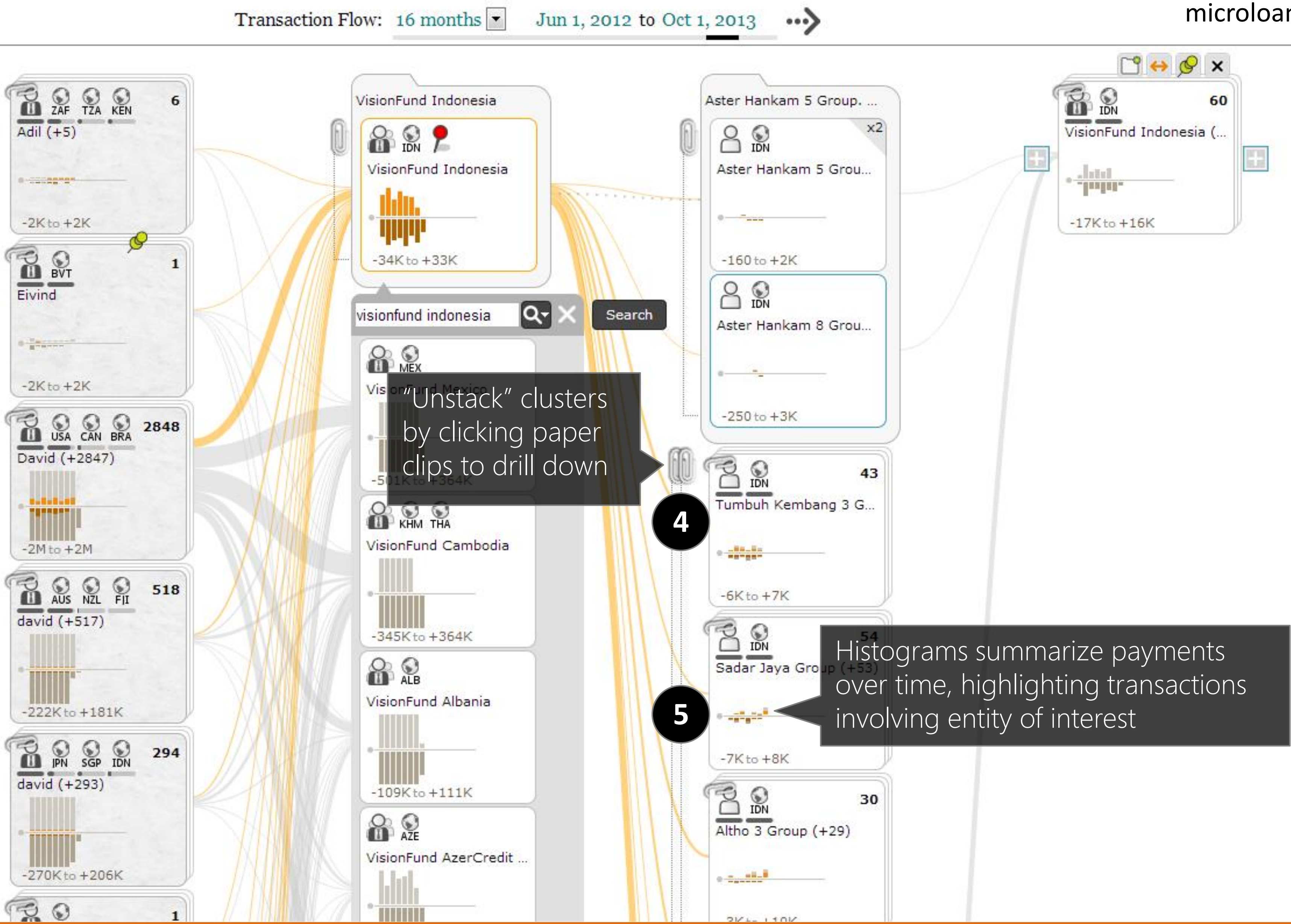
KIVA microloans



oculus

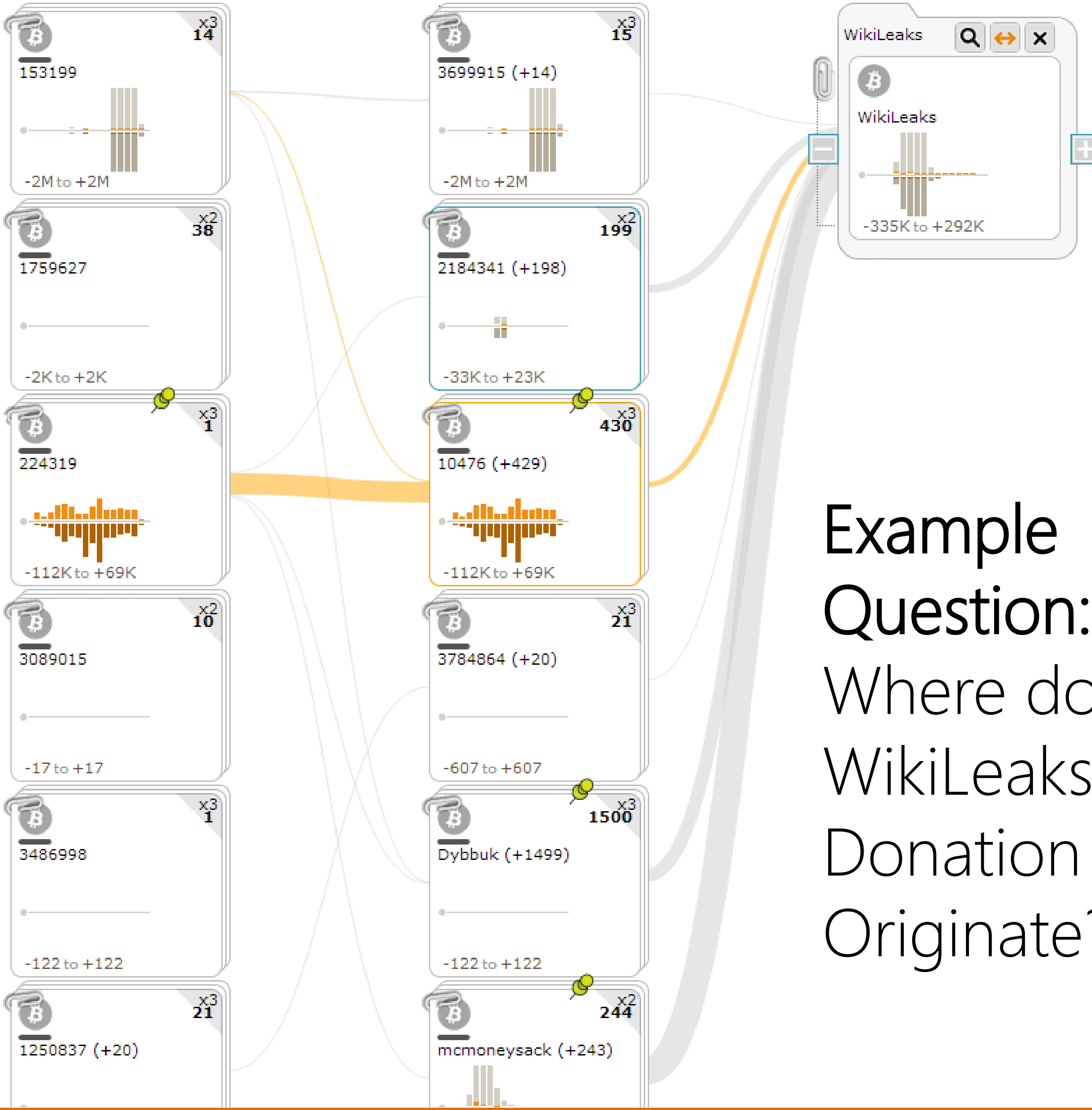
Example: Kiva Partner Transactions

microloans



Example II: Bitcoin Transactions

bitcoin
virtual currency

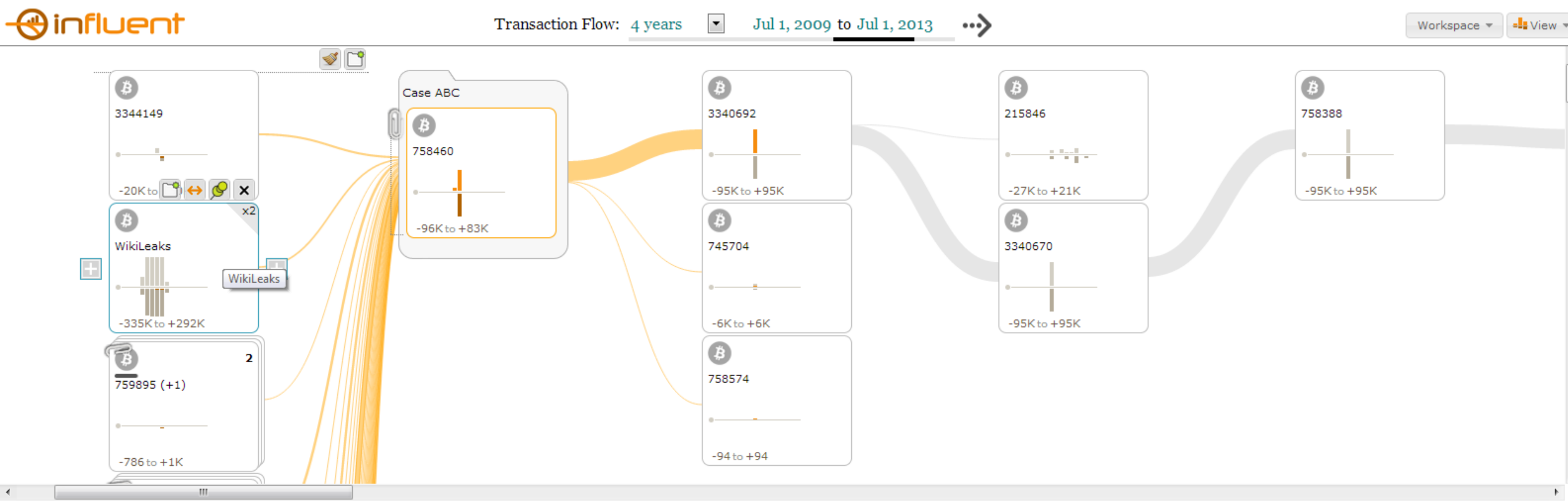


Example Question: Where do WikiLeaks Donation Funds Originate?

oculus

Example II: Bitcoin Transactions

bitcoin
virtual currency



Details

WikiLeaks

uid:	5104
transaction count:	99216
avg transaction amount:	34
max transaction amount:	50000
total transaction amount:	1906758
first transaction date:	26 Apr 2010
last transaction date:	10 Apr 2013
degree:	82523
userTag:	WikiLeaks
LABEL:	WikiLeaks
TYPE:	ACCOUNT

bitcoin

#	Date	Comment	Inflowing	Outflowing
1	2011-03-26	2763.68 BTC to 758460		2,363.50
2	2011-04-16	1925.92 BTC to 758460		2,022.02
3	2011-04-16	1000.0 BTC to 758460		1,049.90
4	2011-04-17	100.0 BTC to 758460		111.23
5	2011-04-18	2003.2 BTC to 758460		2,327.70
6	2011-04-18	217.0 BTC to 758460		252.15

Showing 1 to 6 of 6 entries Previous Next

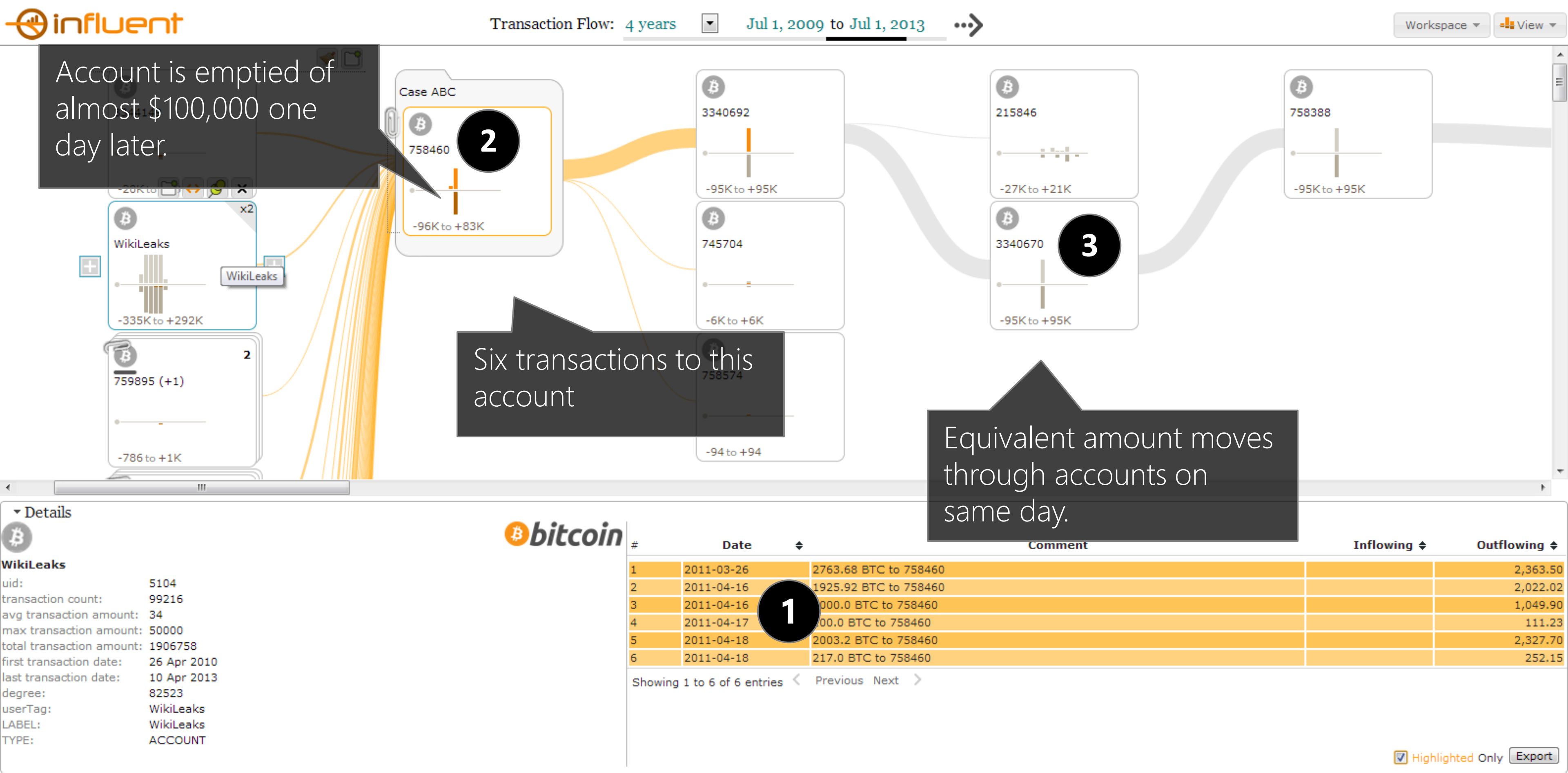
Highlighted Only Export

...and where do they go?

oculus

Example: Bitcoin Transactions

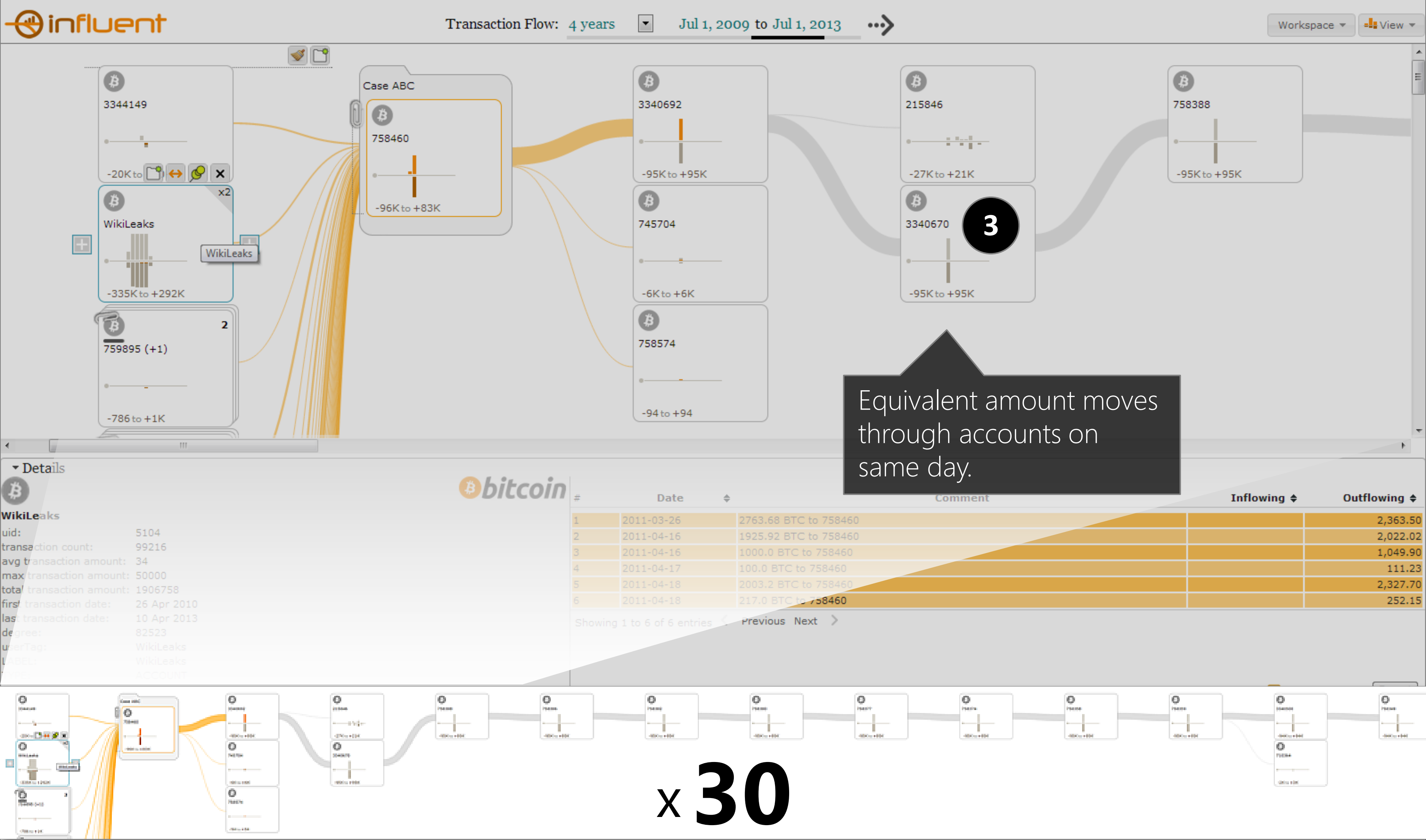
bitcoin
virtual currency



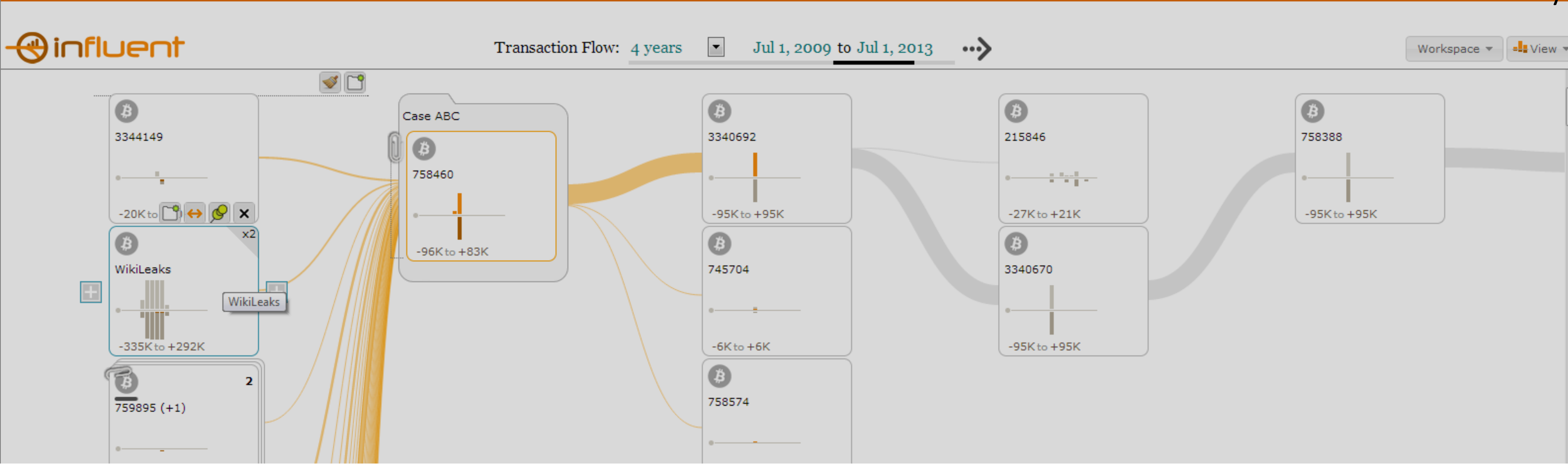
#	Date	Comment	Inflowing	Outflowing
1	2011-03-26	2763.68 BTC to 758460		2,363.50
2	2011-04-16	1925.92 BTC to 758460		2,022.02
3	2011-04-16	000.0 BTC to 758460		1,049.90
4	2011-04-17	00.0 BTC to 758460		111.23
5	2011-04-18	2003.2 BTC to 758460		2,327.70
6	2011-04-18	217.0 BTC to 758460		252.15

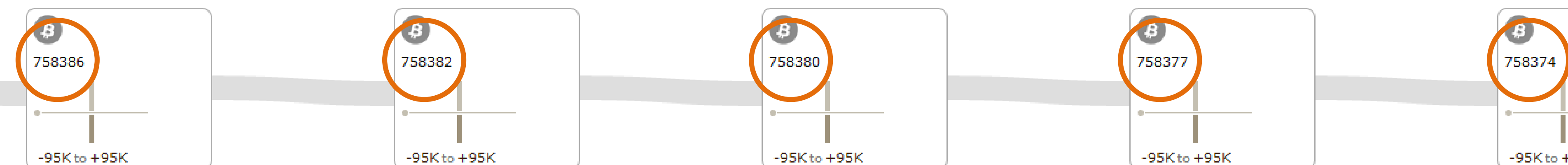
Example: Bitcoin Transactions

bitcoin
virtual currency

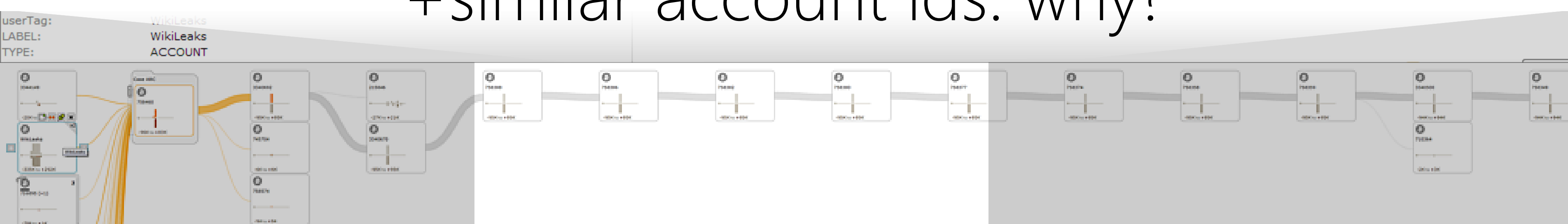


Example: Bitcoin Transactions





+similar account ids. why?



Money Flow Graph Summary

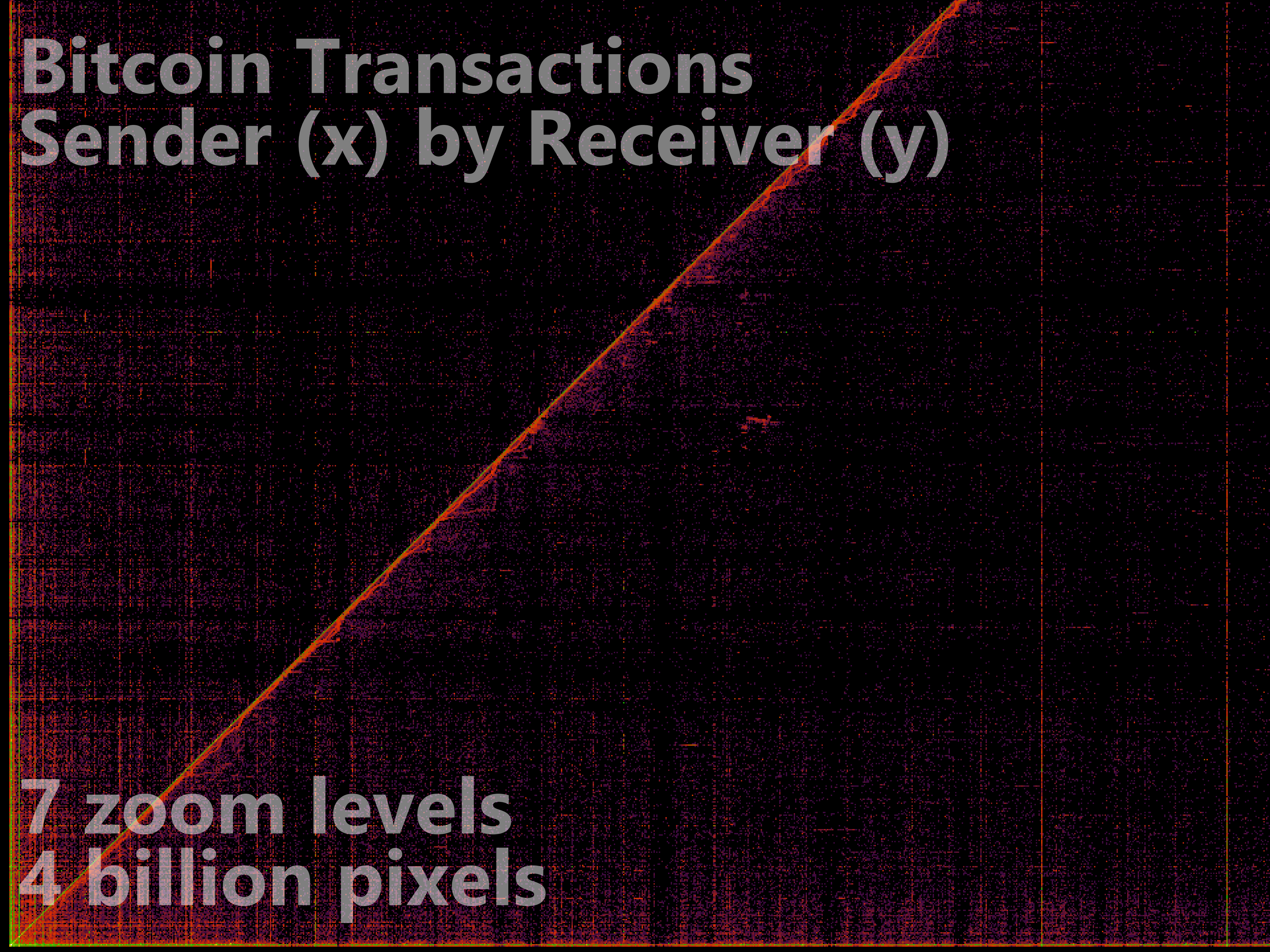
1. Focus on the flow
   + search to start
   + expand

2. Analytics to handle complexity e.g.
   + clustering
   + similarity

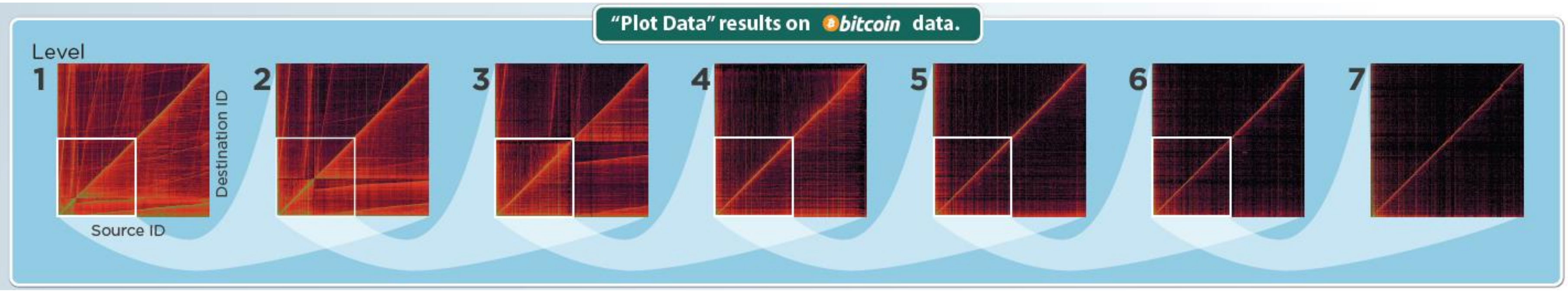
Making Sense of Big Graphs #4

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

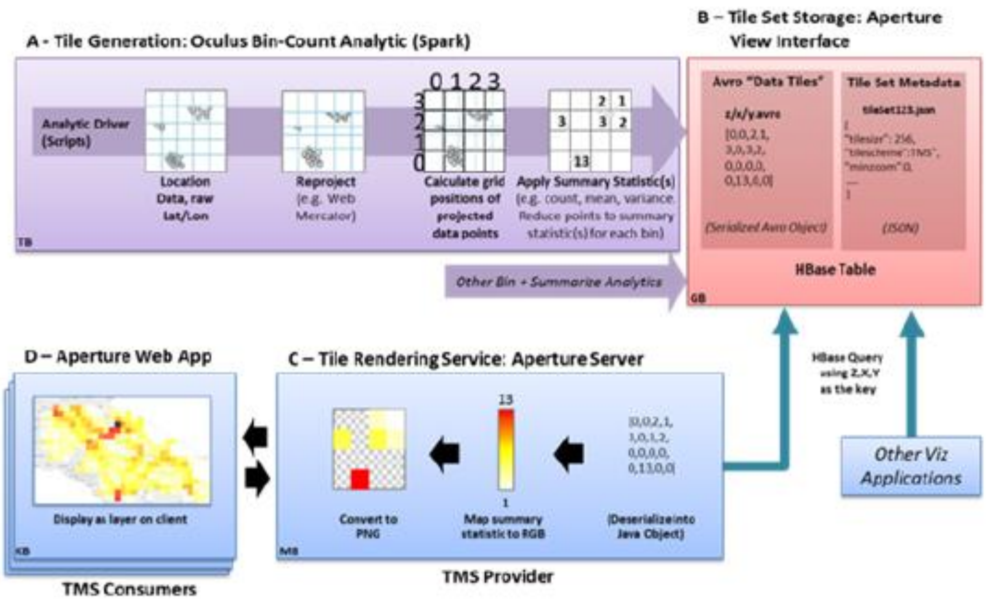
Aperture Tile Based Visual Analytics



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

Using:

[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.

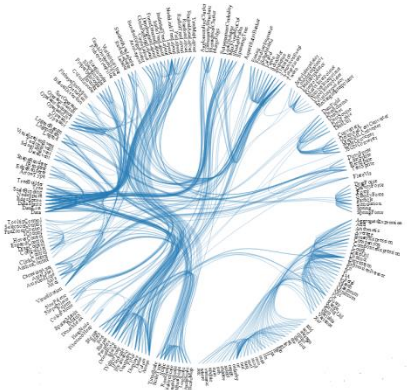
Matrix Graph Summary

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

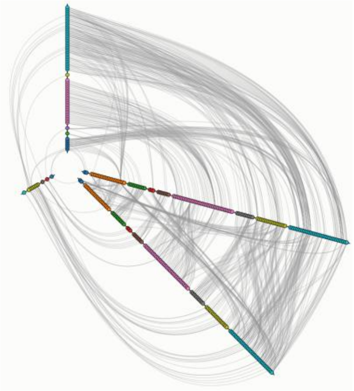
Many more techniques exist

Edge Bundling



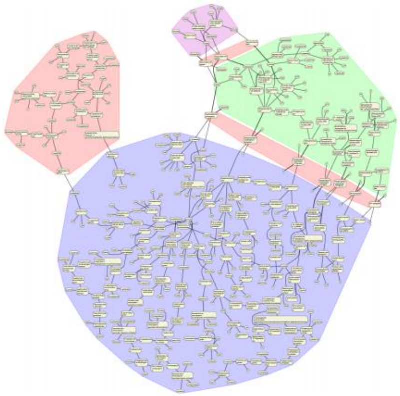
http://mbostock.github.io/d3/talk/20111116/bundle.html

Hive Plot



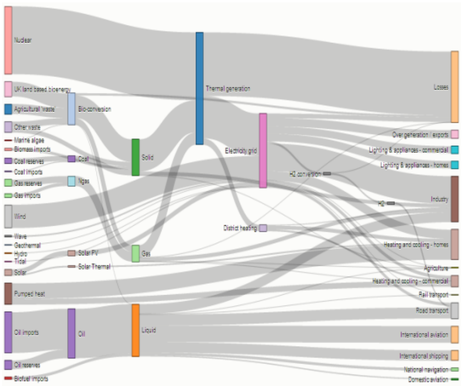
http://bost.ocks.org/mike/hive/
http://egweb.bcgsc.ca/

Hierarchical/Euler



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

Sankey Diagrams



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

Circos

http://circos.ca/

Where next?

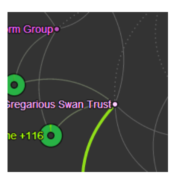
Want to point and click through some graphs?

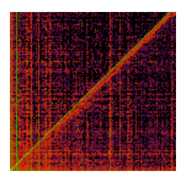
- Gephi (for node and link graphs) www.gephi.org
- Cytoscape (another for node and link graphs) www.cytoscape.org/
- Excel (for matrix graphs)

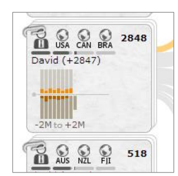
Want to program using some toolkits?

- D3.js d3js.org

- Coming soon to github.com/oculusinfo:



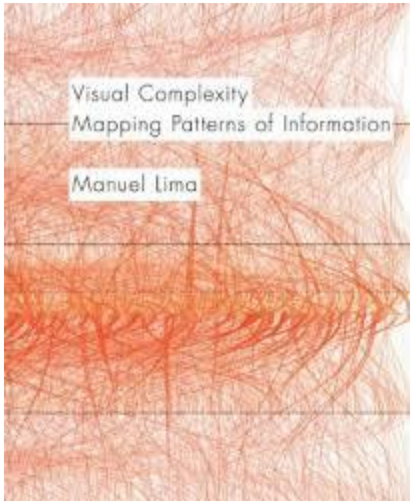




Some nice reference books

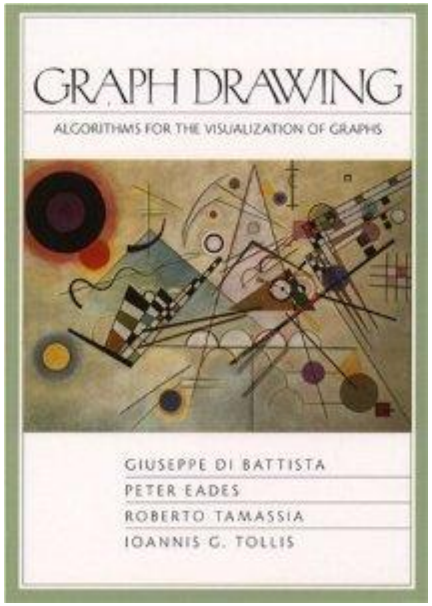
See some really nice looking graph **pictures** for inspiration?

→ **[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)



oculus

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