Centrifuge White Paper
An Alternative to Producing Link Analysis Graphs for Large Datasets
August 2014

This paper describes a technique for creating a useful visualization based upon a set of 800,000 Twitter follower records. Centrifuge can create a link analysis visualization of this data, but such a visualization will typically be difficult to interpret due to the intrinsic limits of displaying this much information. In such a case, there are alternate ways to glean important insights from the data.

Background

When a user on Twitter “follows” another user, this reflects some degree of connection between those two users because it means that the follower wishes to track what the followed-user is saying. (It doesn’t mean that the follower is sympathetic -- it could be the follower disagrees with the other user, but wishes to stay apprised of what he’s saying.)

The collection of this data can be represented as a directed graph where graph nodes are users and graph edges represent the follow relationship.

There are scenarios in which people are interested in analyzing and better understanding who follows whom on Twitter. This could include determining where the “clusters” of followers lie, the overlaps between clusters, etc.

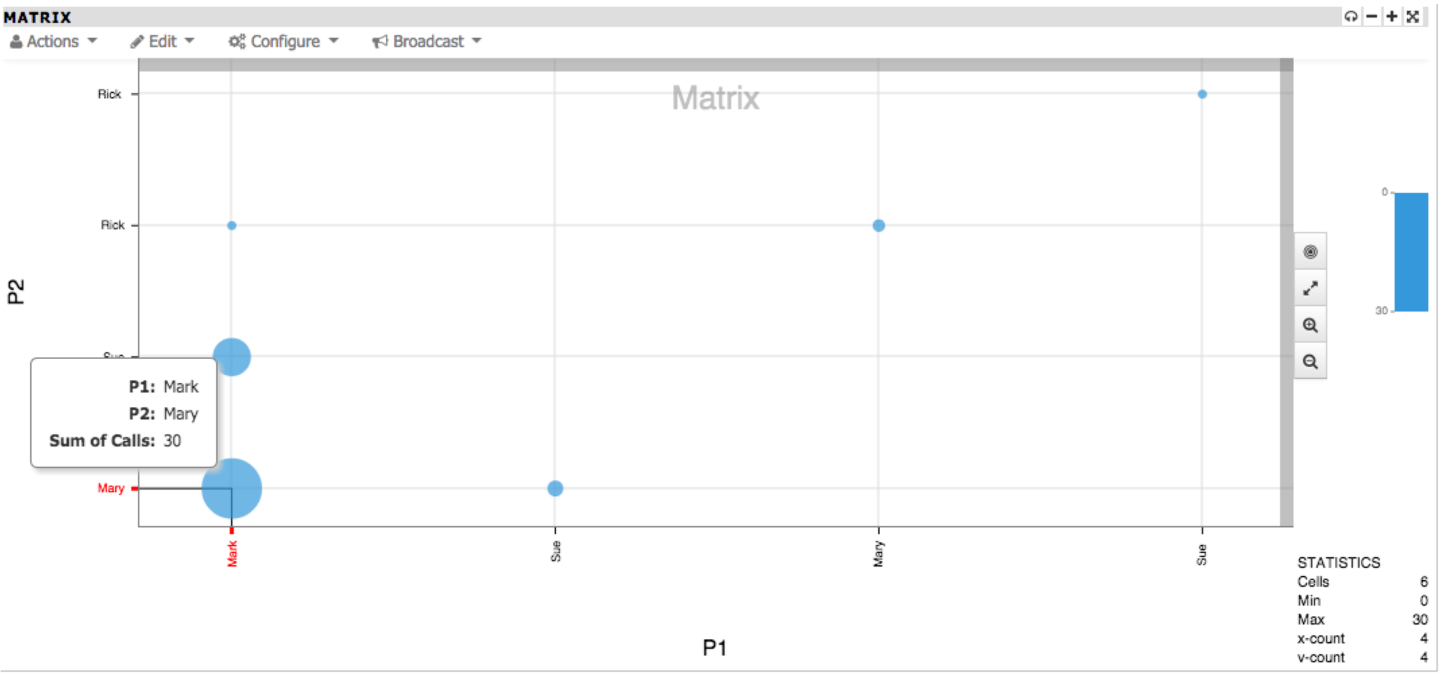
Centrifuge does an excellent job of rapidly computing and displaying a link analysis diagram, even for large sets of data exceeding 100,000 records. But link analysis diagrams with this much data are hard to interpret, in any tool. (Imagine making sense of a photograph of 100,000 people.)

This paper describes an alternate visualization technique in which we pre-process the data to make it amenable to display in what is often called a *bubble matrix.* The bubble matrix is similar to what Excel calls a Bubble Chart, however, in the bubble matrix the labels are the same on both axes (but may not be in the same sequence.) Where the labels intersect, a circle (bubble) is drawn, and the bubble’s diameter is proportional to some value intrinsic to the two dimensions on the axes.

Since this explanation of a bubble matrix probably makes sense only to this author, here’s an example involving phone call data between several individuals.

Person1	**Person2**	**Number of calls**
Mary	Rick	2
Sue	Mary	4
Sue	Rick	0
Mark	Sue	17
Mark	Rick	0
Mark	Mary	30

The resulting bubble matrix of this data looks like this:



P1: Mark
P2: Mary
Sum of Calls: 30

STATISTICS	
Cells	6
Min	0
Max	30
x-count	4
v-count	4

The two large bubbles represent the Mark/Mary and the Mark/Sue relationships. Clearly, those are the two strongest relationships in the data. Centrifuge clusters the larger bubbles in the lower left side of the chart. When the user hovers over a bubble, a tooltip appears to show the two dimensions, and the total.

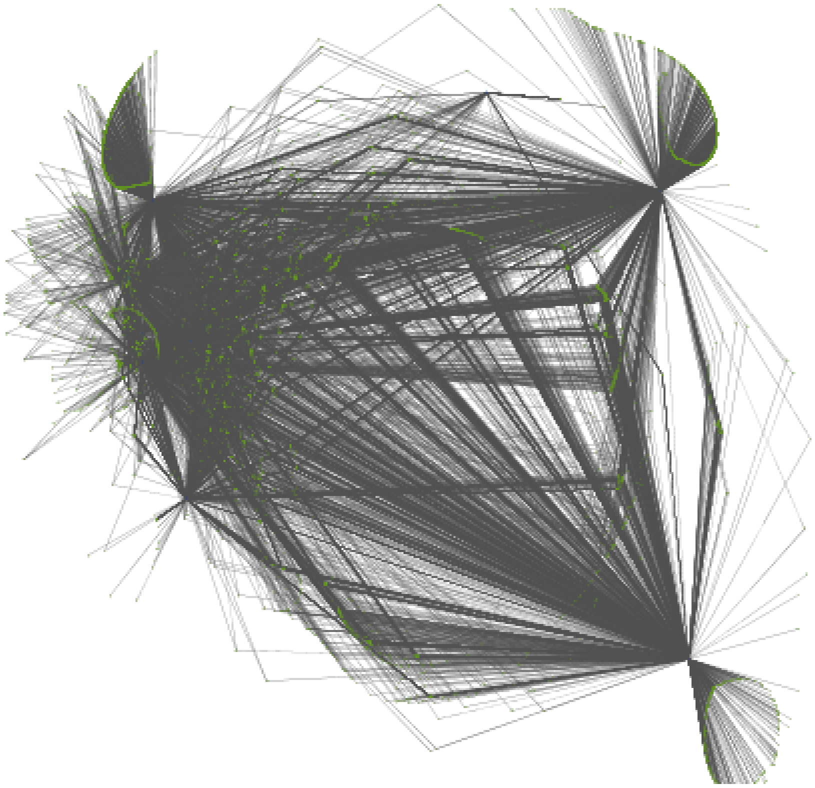
With small datasets like the example above, the value of this visualization could be questioned -- where it shines is with larger datasets.

Back to our Twitter dataset, which is 800,000 rows of data that looks like this:

abuhila,larrys1996
abuhila,Ricklovesmary
abuhila,june32
justkickit,june32
justkickit,redfin
justkickit,JackOfAllTrades
etc.

Each row represents a follower relationship: e.g. larrys1996 is following abuhila. The dataset has 800,000 rows, but only for 28 users (the values in the first column). Thus all of these users are VIP's, they have lots of followers.

The link analysis graph for *only 100,000* rows, a little over 10% of this data, looks like this:



This is an accurate depiction of the data subset, but it's hard to work with, and again it's just 12% of the data we wish to analyze. Filtering features in Centrifuge allow winnowing the data down, and other layouts can provide insights, but clearly this is going to be hard to work with because of its size.

Consequently, we need other visualization alternatives.

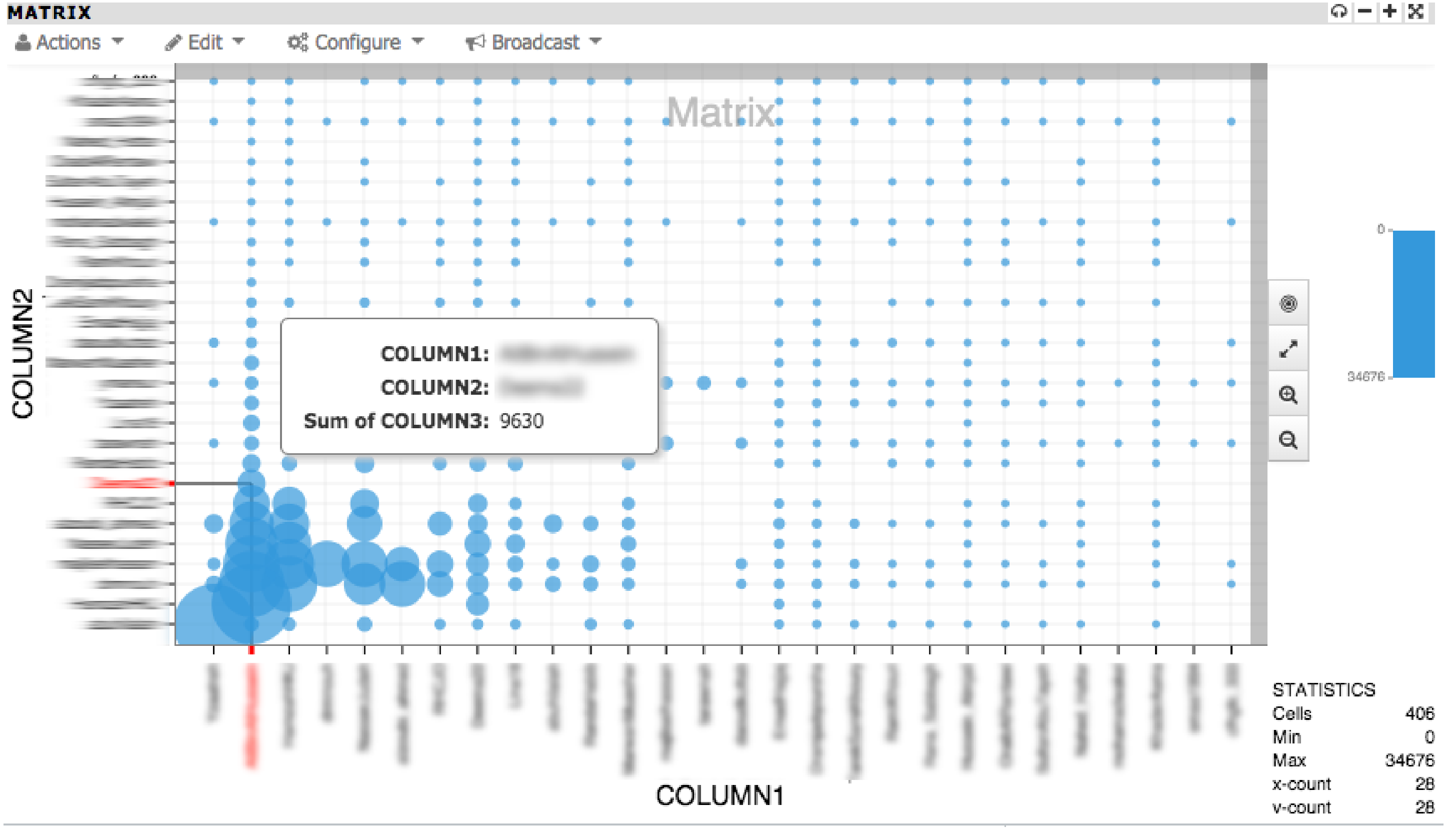
One possibility is to consider how the 28 Twitter users of interest relate to one another. The approach is as follows: compare the follower list of each of the 28 users to the list of every other user, and compute how many followers each has in common.[1]

Looking at the sample data above, justkickit and abuhila have 1 follower in common, so there would be an entry:

justkickit,abuhila,1

[1] Centrifuge professional services personnel wrote software to derive this data. Contact support@centrifugesystems.com to get a copy of this software.

All in all there are 406 pairs from 28 users. We can easily render a bubble matrix with 28 values on each axis. Here's the visualization:



COLUMN1: [redacted]
COLUMN2: [redacted]
Sum of COLUMN3: 9630

STATISTICS

Cells	406
Min	0
Max	34676
x-count	28
y-count	28

The largest bubbles represent the pairs having the most followers in common. Among other things, this tells us that two users with a large intersection bubble ***may*** have similar messages.

We could also pursue creating a link analysis chart for the derived data, but we'll leave that for the next white paper.

Note: This approach is not intended to be the most rigorous mathematical approach to conducting such analysis, rather, it is meant as practical guidance and inspiration for those having to interpret datasets such as this. For example, one refinement to this technique would include a normalization step in which the number of common followers would be divided by the total number of followers for any given pair, in order to equalize the numbers for larger sets. In other words, having 50 common followers would be less significant for two users having a million followers each, than for two users have 100 followers each.