Visualizing Networks

Jon-Michael Deldin

Fall 2011

Contents

1 Introduction

For this project, I am using two data sets, "Zachary's Karate Club" and the Enron email corpus to explore graphing social networks. Note: My graphs are full-page and located at the end.

2 Examples

One example of a good network visual is the Euro Crisis image shown in Figure 1. This figure displays the interconnected debt crisis in a useful way because it weights each edge according to the debt owed, similar to Joseph Minard's French wine export (Figure 2). Additionally, edges and nodes are

annotated to give the viewer much more information than a simplified network graph. These features and the instructions at the top make this figure self-documenting and very easy to use.

An example of a poor network visualization is the image shown in Figure 3 of world-class universities on the Internet. This figure is supposed to show the Yahoo! search engine rankings of by scaling the orbs, and it also tries to show universities in the same country with edges. Supposedly it also shows geographical relationships, but between the dark gray lines, overlapping orbs, and labeled nodes, one cannot arrive at any conclusions. It is a poor graph because it is doing way too much in one figure for a questionable hypothesis.

3 Zachary's Karate Club

"Zachary's Karate Club" is a small social network from W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* **33**, 452-473 (1977). The network is comprised of 34 nodes and 78 edges, where the nodes represent club members, and the edges represent friendships outside of the karate club. The data is of the network before the club split into separate groups over paying their instructor more money for lessons.

Based on my knowledge of the 1970's from movies and family stories, it seems like people actually socialized in and outside of clubs. Thus, I hypothesized that most of the members would be friends (i.e., more than one edge per node).

To support my hypothesis, I created an undirected, degree-sorted circular layout graph in Cytoscape. The figure clearly shows the multiple friendships between members. This image is consistent with my good example and the class discussions because it is self-documenting by describing the data, data source, nodes, and edges; it uses a reduced color palette; and it provides insight into the structure of the network with a bar chart of the nodes and edges. Additionally, I tried to implement the Golden Ratio, but I ended up using the rule of thirds for the text. Finally, I produced the bar chart using matplotlib and assembled the graphs and text in Adobe Illustrator.

4 Enron

I investigated a number of hypotheses related to the Enron email corpus available at `http://www.cs.cmu.edu/~enron/`.

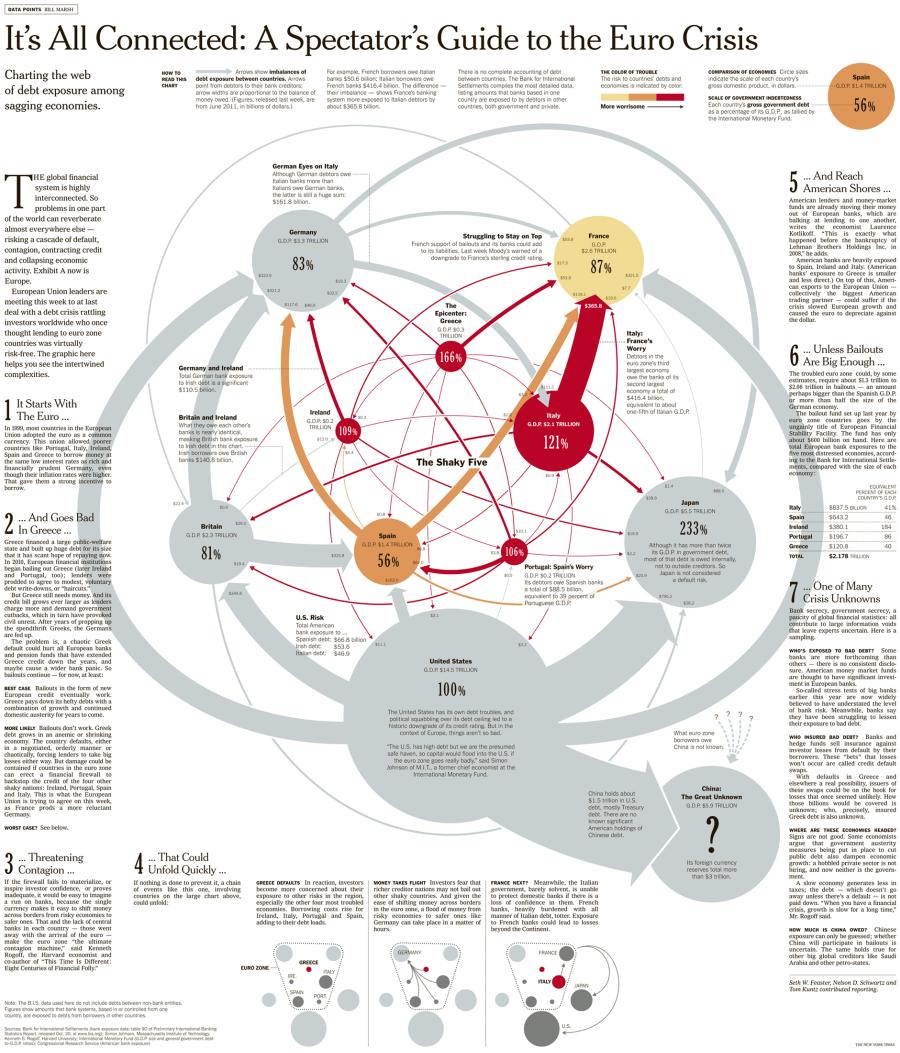


Figure 1: Illustration of the 2011 Euro Crisis by The New York Times from `http://www.nytimes.com/imagepages/2011/10/22/opinion/20111023_DATAPOINTS.html`.

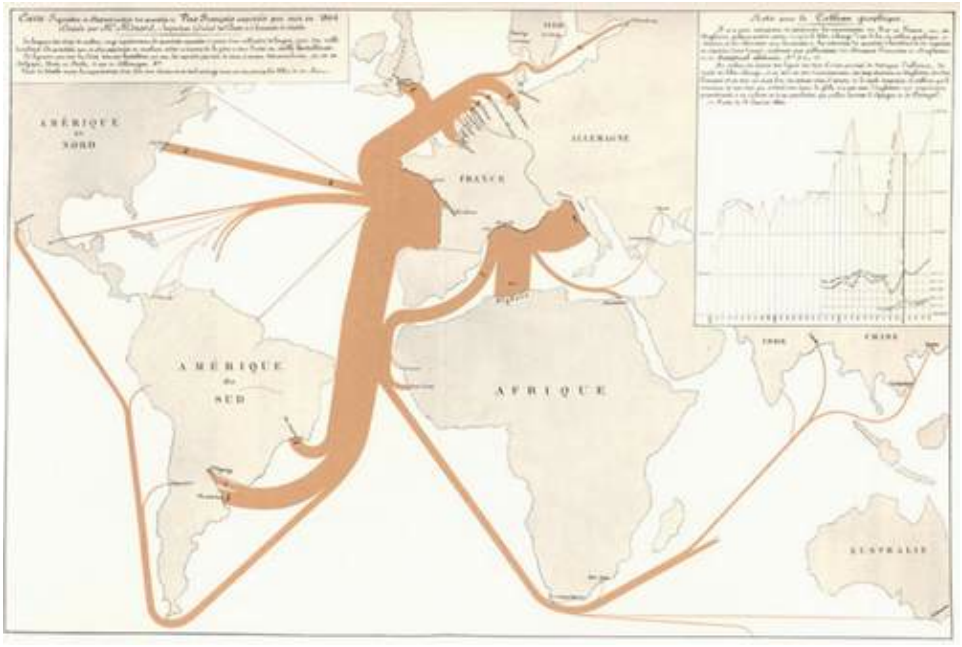


Figure 2: Joseph Minard's visualization of French wine exports from `http://hci.caece.net/hci-wiki/CoursePresentation2007/FlowMapLayout`.

The data set consists of 2.5 GB of plain-text emails – 517,440 total email messages, of which 126,020 are sent messages. The messages span from 1998-10-30 to 2002-07-12. Unfortunately, it seems the key people in the scandal (e.g., Kenneth Lay and Jeffrey Skilling) either communicated very little over email or destroyed the messages (while shredding their records, of course). However, with over 500k messages, there is still valuable network information.

The `maildir` format was unwieldy, so I wrote a data parser and loader in Ruby to extract the following header fields: `From:`, `To:`, `Cc:`, `X-cc:`, `Subject:`, and `Date:`. Once extracted, I inserted the fields into a Postgres database (afterwards, I found a MySQL dump file of the entire data set at `http://bailando.sims.berkeley.edu/enron_email.html`).

4.1 Jeff Dasovich

Jeff Dasovich was the government relations executive for Enron. I stumbled across him when initially exploring the data set, looking for "spammers" (Table 1), and found he'd sent the most emails out of anyone. After researching

him, I discovered he was aware of Enron's role in the California energy crisis of 2000-2001[1]. Based on that and the number of emails he sent, I then hypothesized that he must be communicating with some of those involved in the scandal.

Table 1: This shows employees who sent the most email at Enron. The number of addresses reached is the sum of the number of unique emails in the `To`, `Cc`, and `X-cc` fields.

Email	Addresses Reached
`jeff.dasovich@enron.com`	34765
`kay.mann@enron.com`	15025
`tana.jones@enron.com`	14726
`chris.germany@enron.com`	14226
`vince.kaminski@enron.com`	10010

To support my hypothesis, I created a poster of Dasovich's email profile. I included a narrative, a matplotlib time series of his sent emails, and his top 15 recipients as a network. I plotted the network using a directed, circular layout in Cytoscape. I then weighted each edge by the number of emails sent to a recipient (node) and reflected this weight using the line-width. The resulting figure supports my hypothesis because it shows Dasovich was in contact with Tim Belden, who is considered the mastermind of California's energy crisis.

The image is consistent with the class discussions and my example for the same reasons my karate image was, but this one adds more value by showing the edge count on the nodes and integrating a time series from matplotlib into the narrative.

4.2 Structure of Enron

For my final figure, I hypothesized that clusters would be present when plotting Enron's email connections. I created a plot in Cytoscape using a degree-grouping layout to create the conical image at the end of this document. This supports my hypothesis because you can see large rings where people only received emails from one source and small rings that had multiple sources. The flow of email is self-organizing.

[1] `http://money.cnn.com/2006/04/04/news/newsmakers/enron_blog_fortune/index.htm`

This image is consistent with the class discussion because it is very data-heavy and provides insight into the structure of the graph through the narrative. However, the amount of data comes at a price: many clusters are obscured, and moreover, one cannot easily determine the origins of many edges. (The web version[2] may be easier to visualize the clusters – the transparency didn't preserve well in print.)

To refute my hypothesis, I used the same data set and created a circular layout. Now, none of the clusters are preserved, which does not support my hypothesis. This is one of the drawbacks of Cytoscape and most graphing tools – the data is bound to the layout algorithm. In researching tools, I came across the "hive plot"[3], which is a new, unique way of visualizing complex networks. Nodes are constrained to axes, and positioning of the nodes is determined by network structure and not the layout engine. Unfortunately, the paper describing it is in press, and the documentation for the `linnet` tool is absent. Nevertheless, it is a promising technique that I hope to utilize in the future.

5 Conclusion

Overall, I was able to support all of my hypotheses and produce graphs of sparse to dense networks. This was the most challenging and rewarding assignment yet, requiring countless hours of tweaking parsers to pixels for a final product.

[2] `http://jmdeldin.com/dv/graph/fig/enron_struct_trans.png`
[3] `http://hiveplot.com/`

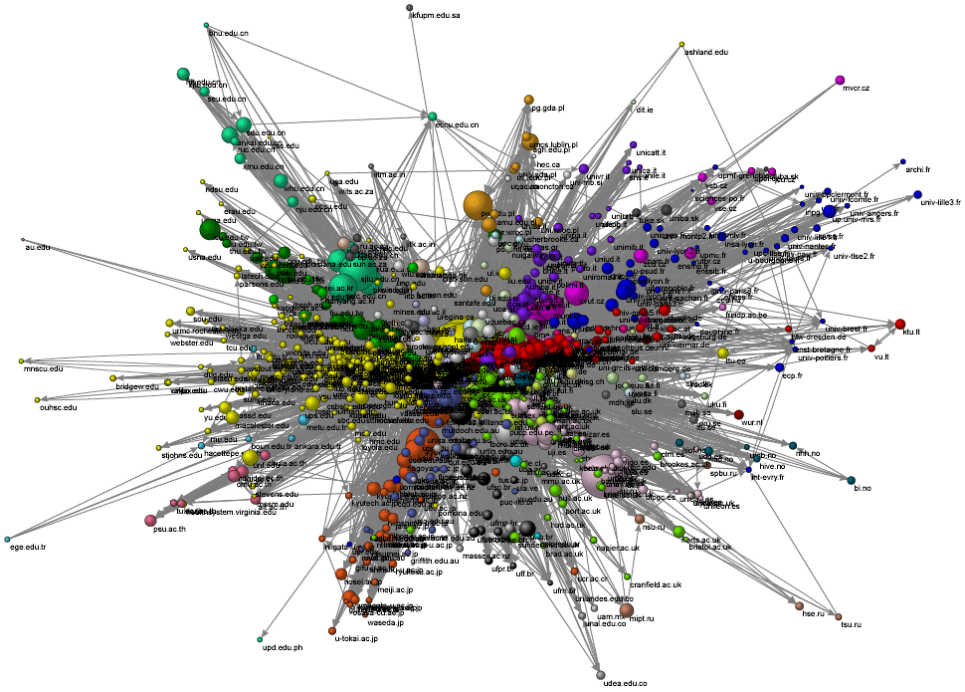
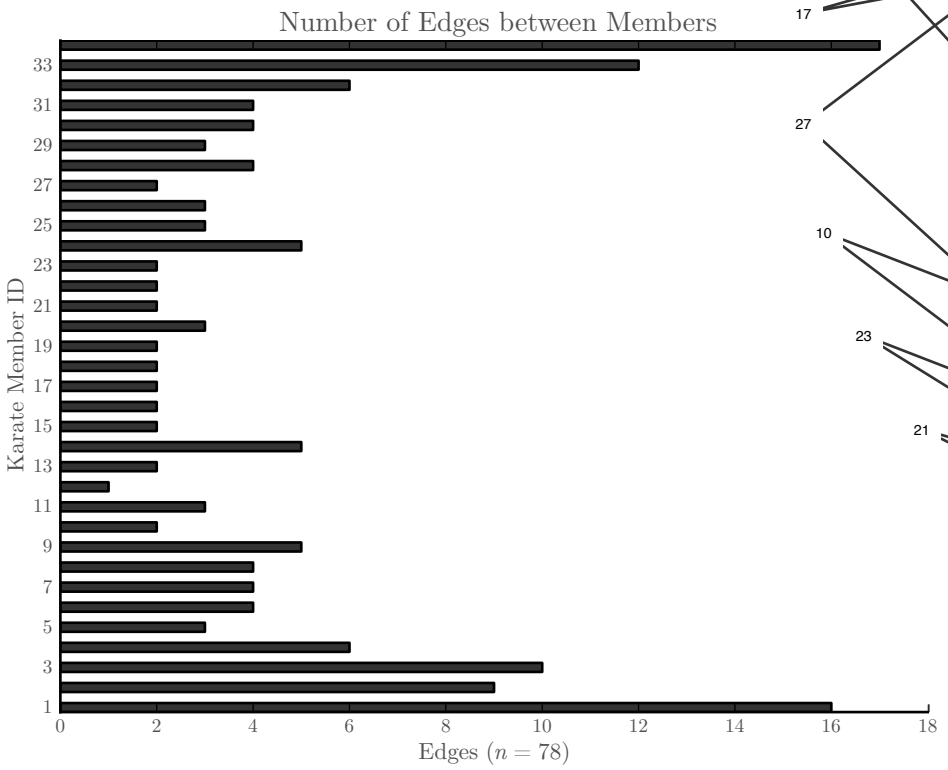


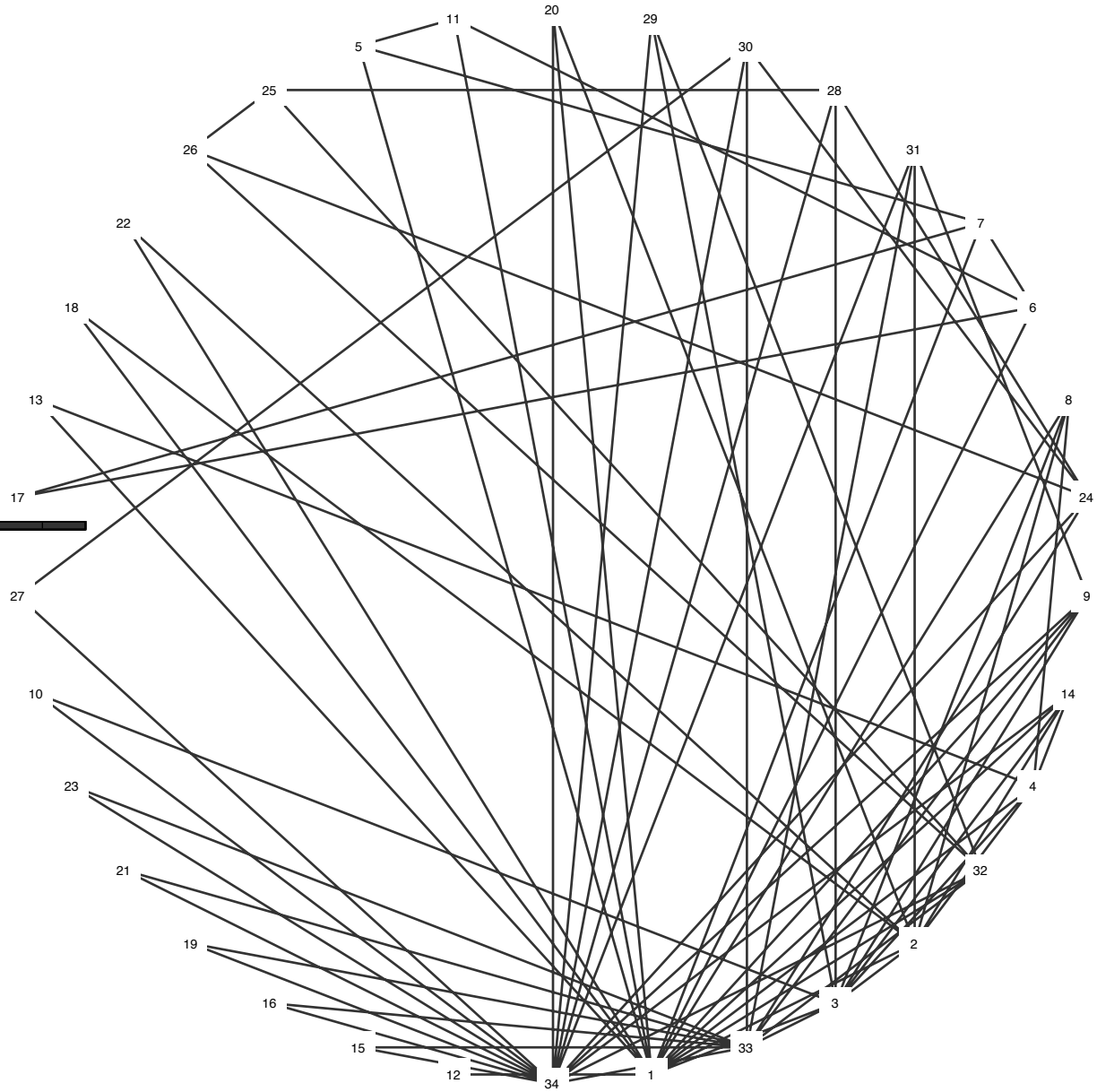
Figure 3: A network graph of world class universities on the Internet from `http://internetlab.cindoc.csic.es/cv/11/world_map/map.html`.

karate club relationships

"Zachary's Karate Club" is a three-year study of 34 members of a karate club before the group split over karate lesson prices. The data originally appeared in W. W. Zachary, An information flow model for conflict and fission in small groups, *Journal of Anthropological Research* 33, 452-473 (1977).

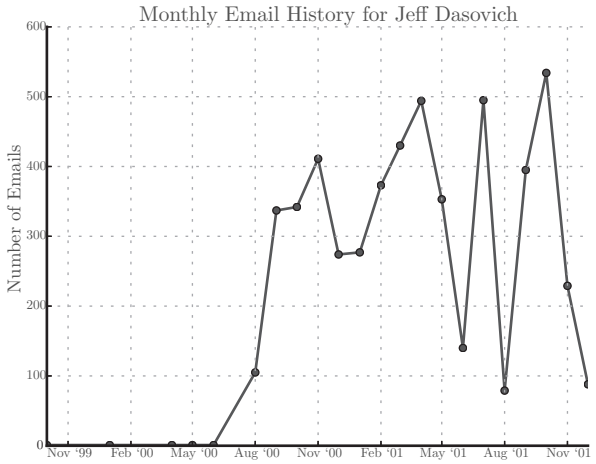
We can represent the karate club with a degree-sorted circular layout. Nodes represent individual club members (anonymized with an ID), and edges represent friendships outside of the karate club.





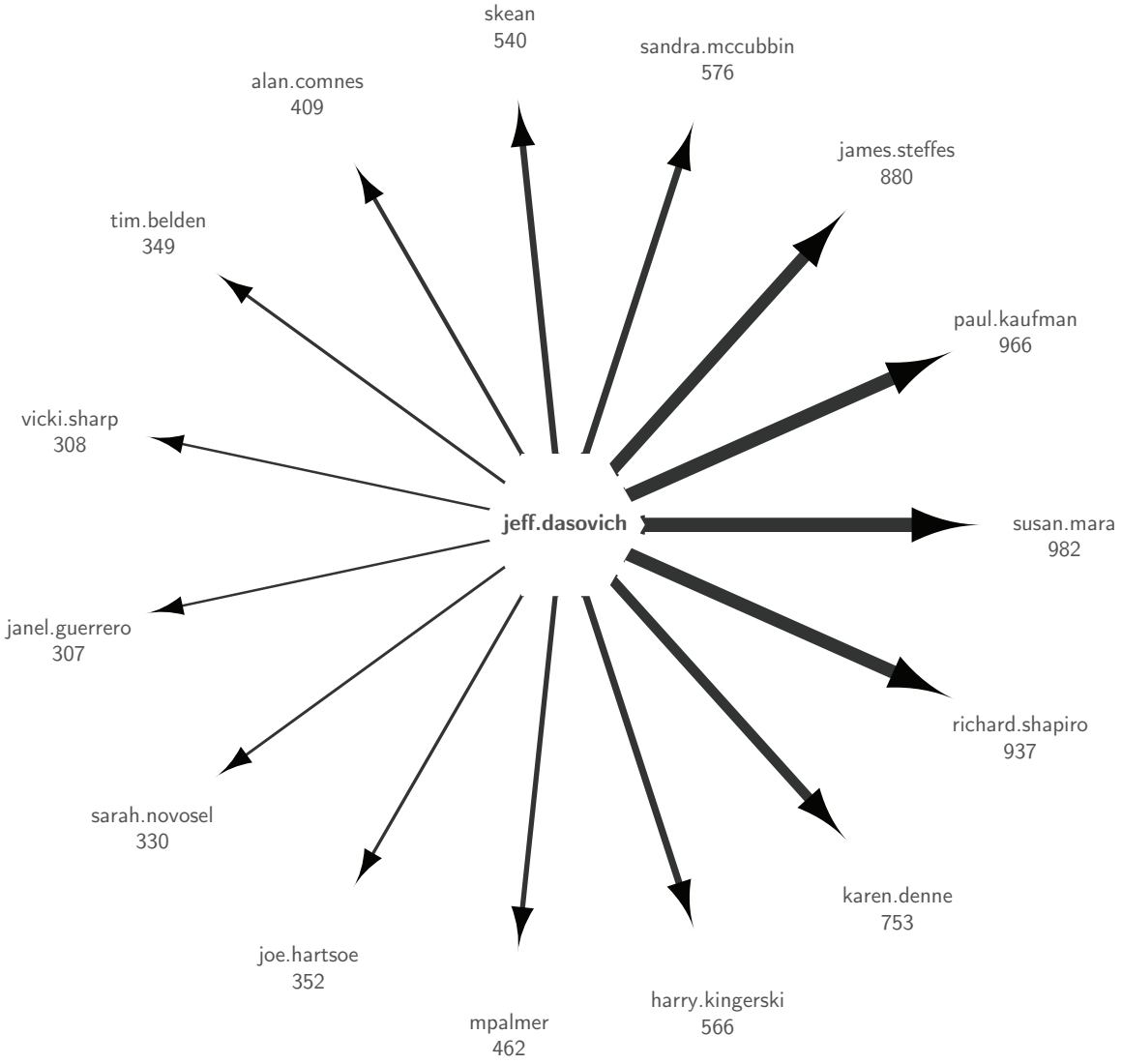
email profile: jeff dasovich

Jeff Dasovich, former government relations executive for Enron, may not have been a major player in the Enron scandal, but electronically, he had a significant impact. From 1999-10 to 2001-12, he sent 5,361 emails, but he CC'd multiple people on each message, resulting in 34,765 messages. Who was he emailing the most?



Plotting all of Dasovich's communications would result in an enormous hairball, but we can focus on his top 15 recipients. To the right, each node is an `@enron.com` address, and each edge is a weighted representation of the number of emails sent to that address by Dasovich. We can see he was in frequent contact with Richard Shapiro (VP of regulatory affairs), and Tim Belden (head of trading in Enron Energy Services and mastermind of California's 2000–1 energy crisis).

The data set is derived from the scrubbed collection of Enron email messages maintained by W. Cohen at `http://www.cs.cmu.edu/~enron/`.



For a clearer image, view this at http://jmdeldin.com/dv/graph/fig/enron_struct_trans.png

structure of enron

Plotting Enron's employee emails in a network graph reveals a multitude of clusters inside Enron (right). The largest cluster corresponds to those employees who received emails from one source, whereas the smaller clusters are comprised of people receiving emails from multiple sources. In effect, we can see the communication groups of the company through the 6,336 nodes and 17,752 edges.

Unfortunately, plotting this data is highly dependent on the layout algorithm used – changing from grouping by degree to a circular layout (below) loses the clusters and makes identifying important nodes extremely difficult.

The data set is derived from the scrubbed collection of Enron email messages maintained by W. Cohen at `http://www.cs.cmu.edu/~enron/`. Only the year 2001 was plotted to minimize the "hairball plot" effect.