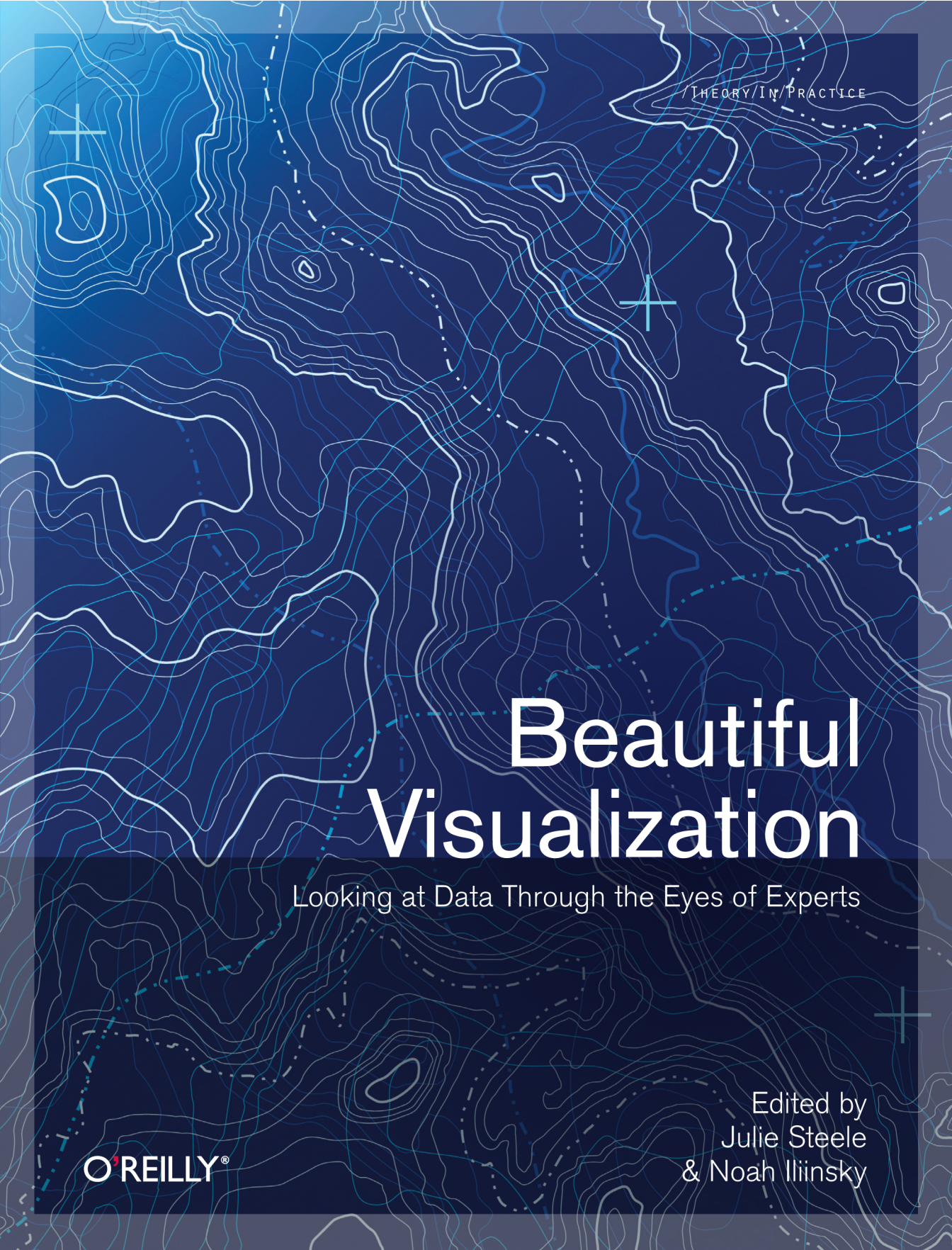
/Theory/In/Practice
Beautiful Visualization
Looking at Data Through the Eyes of Experts
Edited by
Julie Steele
& Noah Iliinsky
O'REILLY®

CHAPTER SEVEN

Your Choices Reveal Who You Are: Mining and Visualizing Social Patterns

Valdis Krebs

DATA MINING AND DATA VISUALIZATION GO HAND IN HAND. Finding complex patterns in data and making them visible for further interpretation utilizes the power of computers, along with the power of the human mind. Used properly, this is a great combination, enabling efficient and sophisticated data crunching and pattern recognition.

In this chapter, we will explore several datasets that reveal interesting insights into the human behaviors behind them. Patterns formed by event attendance and object selection will give us clues into the thinking and behavior of the humans attending the events and choosing the objects. Often, our simple behaviors and choices can reveal who we are, and whom we are like.

Early Social Graphs

In the 1930s, a group of sociologists and ethnographers did a small "data mining" experiment. They wanted to derive the social structure of a group of women in a small town in the southern United States. They used public data that appeared in the local newspaper. Their dataset was small: 18 women attending 14 different social events.

They wondered: could we figure out the social structure (today we call it a *social graph*) of this group of women? To this end, they posed the following questions:

- Who is a friend of whom?
- Which social circles are they all in?
- Who plays a key role in the social structure?

Identifying network structures normally involves invasive interviews and surveys. Would it be possible to derive network structures by just examining public behaviors? The real question was: *do public choices reveal who you are* and whom you are like?

Being able to see actual connections inside any human system, organization, or community is critical to understanding how groups work and how their members behave. Social network analysis (SNA) is a currently popular set of social science methods used for marketing, improving organizational effectiveness, building economic networks, tracking disease outbreaks, uncovering fraud and corruption, analyzing patterns found in online social networks, and disrupting terrorist networks. SNA techniques can also reveal underlying network structures in the Southern Women dataset, as we will see in a bit.

SNA started as *sociometry* in the early 20th century. Jacob Moreno's drawings of friendship links (or *sociograms*) between students in his school are very popular amongst social science historians, and business scholars point to the famous Hawthorne factory worker studies from earlier in the century and the sketches of work interactions between the "Bank Wiring Room" employees. Friendship ties amongst the Wiring Room employees are illustrated in Figure 7-1.

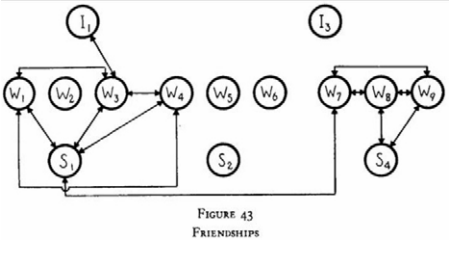

Figure 7-1. *Early 20th-century social graph used in studying workflows amongst employees*

SNA maps a human system as nodes and links. The nodes are usually people, and the links are either relationships between people or flows between people. The links can be directional. When the nodes are of only one type—for example, people, as in the Moreno and Hawthorne studies—it is called *one-mode analysis*.

However, the Southern Women study began as a slightly more complex form of social analysis: two-mode. There were two sets of nodes—people and events—and the links showed which people attended which events. The social graph for the two data modes are shown in Figure 7-2. The women are the blue nodes on the left, while the events that each attended are the green nodes on the right. People are represented by circles, while events are represented by squares.

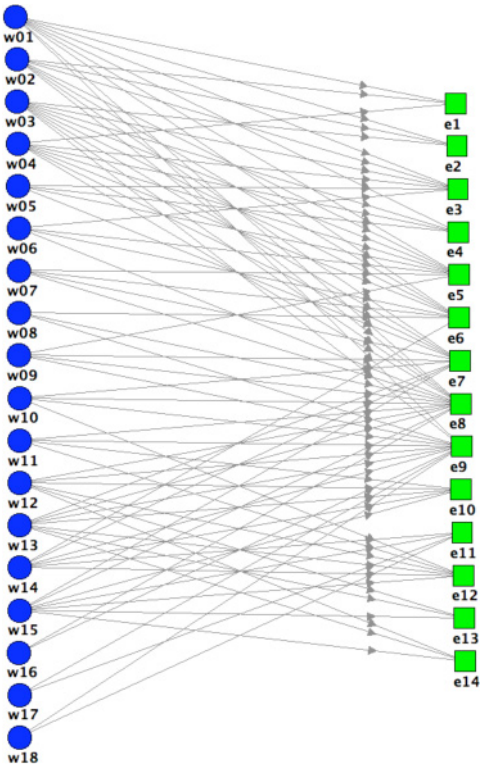

Figure 7-2. *Two-mode view of the Southern Women social event dataset*

This diagram reveals various types of conclusions, such as:

- Woman #3 attended more events than woman #18.
- Event #8 had the most attendees.

Other than these simple observations, the two-mode view does not reveal any obvious patterns, such as the women's social structure or the relationships among the events. To see these deeper insights, we can transform the two-mode data into one-mode data by using a popular social network analysis technique: *transforming nodes to links*. In the first transformation, we'll take the event nodes and view them as links instead:

> Woman X is connected to woman Y as they both attended Event Z.

The more events the women attended together, the stronger their tie is. We can also shift the focus to look at the network of events:

> Event A is connected to event B if they were both attended by the same woman, C.

The more women who attended the same two events, the stronger the connection is between the two events. There are many methods to calculate the link strength when transforming a two-mode network to a one-mode network. In this example, we use the simplest method: adding up the co-occurrences.

The network of events is shown in Figure 7-3. A thicker line reveals a stronger relationship between two events—i.e., that more women attended both events. The SNA software organizes the network according to who is connected to whom using an advanced graph layout algorithm: a node's place in the network is determined by its connections *and* the connections of those connections.

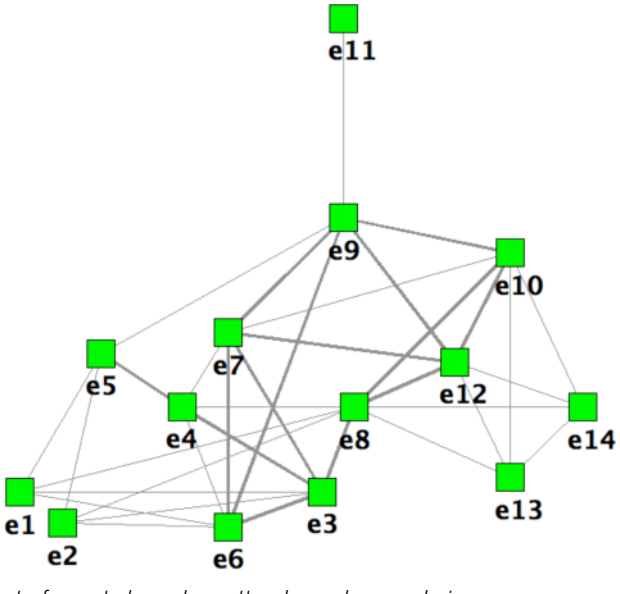

Figure 7-3. *Layout of events based on attendance by people in common*

The center of the graph attracts the better-connected nodes, while the less-connected nodes are pushed toward the periphery. Thus, it is obvious at a glance which events were most important in this social calendar. However, we still do not have a picture of what interests us the most: the emergent social network of the women in this small town. To begin to reveal that network, I used my *gradual inclusion* method, which focuses initially on the strongest ties in the structure and then gradually lowers the membership threshold to reveal weaker ties in the network, allowing more people to connect to whoever is there already. This method usually ignores the very weak ties in the data, dismissing them as *social noise*. In this case, with the small dataset, the dismissal of light connectivity must be done carefully. In a dataset with millions of nodes and millions of choices, adjusting the bar for social noise is usually a less delicate operation.

Using a five-point scale, with 5 indicating the strongest tie between two nodes and 1 indicating the weakest, I started using my gradual inclusion method with *strength = 5* links—in other words, identifying those women who had attended the most events in common. Figure 7-4 reveals the strongest ties based on event attendance.

I immediately saw two clusters form: one with women #1, #2, #3, and #4, and the other with women #12, #13, and #15. I colored the nodes using two different colors to distinguish the membership in each group.

Next, I included the next lower level of ties: *strength = 4* links. This resulted in new members being included in each cluster, but did not reveal any connection between the two clusters. As you can see in Figure 7-5, we still have two distinct groups.

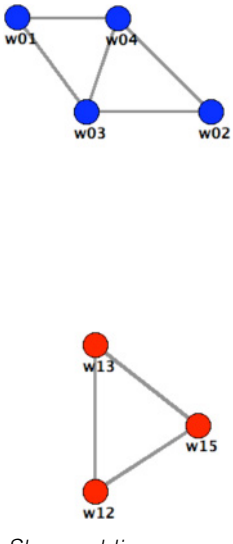

Figure 7-4. *Strongest ties amongst women based on common event attendance*

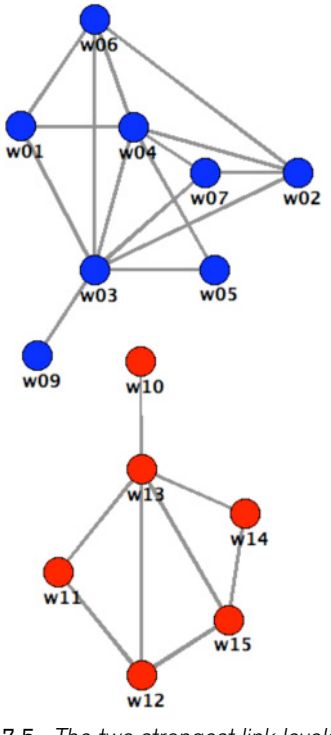

Figure 7-5. *The two strongest link levels between women attending common social events*

Including the *strength = 3* ties revealed bridging between the groups, as Figure 7-6 illustrates. This is common in most social structures: the strongest ties occur *within* a group, while the weaker, less frequent ties occur *between* groups. There were also some weaker ties within each group, indicating that not everyone within a given group has a strong tie to all members of that group.

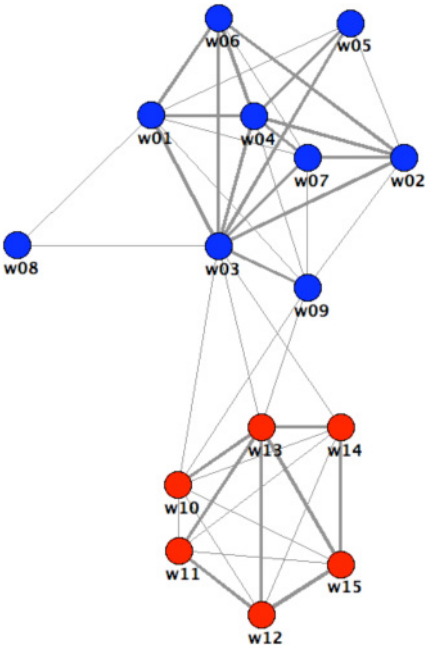

Figure 7-6. *The two groups are bridged with gradual inclusion of weaker ties*

Our social structure is still missing a few nodes: women #16, #17, and #18. They have not met the criteria for attachment in any of the previous waves of inclusion using the gradual inclusion method. Perhaps they are new in town, or are just less social and have attended fewer events, making it more difficult to determine their membership. These three women attach to the network when I lower the threshold to *strength = 2* links. Now all women are attached to the network, while the original two-cluster structure remains. Woman #16 is the only one that does not obviously belong to one cluster or the other; she has equal infrequent ties to both clusters. I therefore classify her as a member of *neither* cluster (not both clusters!) and color her purple. The final emergent social graph is shown in Figure 7-7.

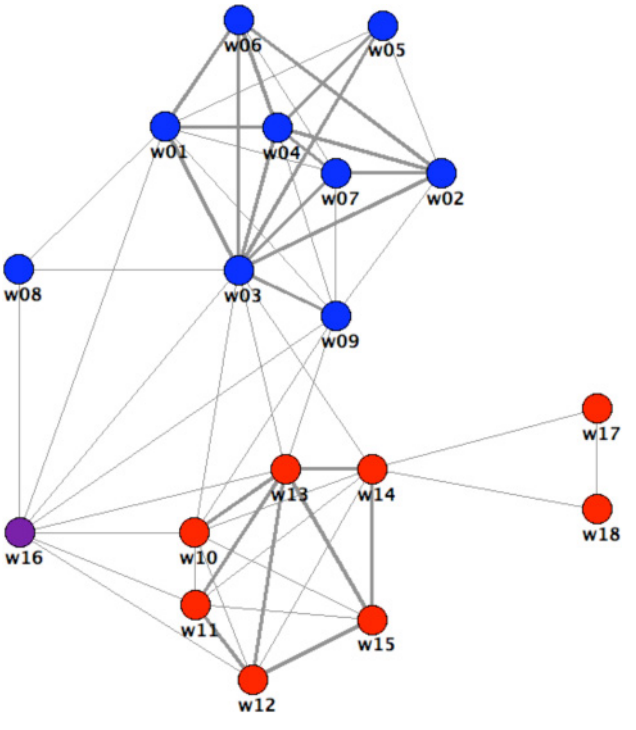

Figure 7-7. *Emergent social graph of women based on common attendance at social events*

All 18 women have now been placed in the social network based on their attendance of local social events. This social network reveals a few interesting things about this small town's social structure:

- Two distinct social clusters exist.
- The clusters are connected. This social overlap reveals some possible commonality in interests and relations between the two clusters.
- Various network roles emerge. Some women are connectors, bridging the two clusters, while others act as internal core members, connecting only to their own groups.

Social graphs like that in Figure 7-7 can be used for marketing purposes or word-of-mouth campaigns. More information can usually be gathered than this simple example provides, but some deductions can nonetheless be drawn from this data:

- Woman #6 will probably not be influenced by what woman #12 does or says.
- Woman #4 probably has the highest internal influence within the blue cluster. She may be the one that reinforces the status quo with everyone in her group.

- Woman #9 in the blue cluster is the *boundary spanner*—the person bridging the two clusters—and probably brings new ideas and opinions into the group. It is good that she has at least one strong tie within the group, to woman #3, who is well connected within the group. People who bring new ideas into a group often need at least one strong, internally well-connected ally.
- Women #16, #17, and #18 may be new in town or may not be "joiners." They have some access to what is happening in the groups, but they may not have access to the real private information in either group because of their weaker connections.

Different data-mining algorithms often produce different results, even with a small dataset such as this one. Over the years, various sociologists and network scientists have re-examined this interesting little dataset, applying their fresh new algorithms to see what patterns emerge. Figure 7-8 shows the results from 21 of the most popular studies. Our results match those of study #13, by Linton Freeman (Freeman 2003): women #1–9 are in one group, women #10–15 and #17–18 are in the other group, and woman #16 belongs to both groups. Freeman was a key player in establishing the field of social network analysis (Freeman 2004) and was especially important in establishing some early network metrics that are still popular today (Freeman 1979).

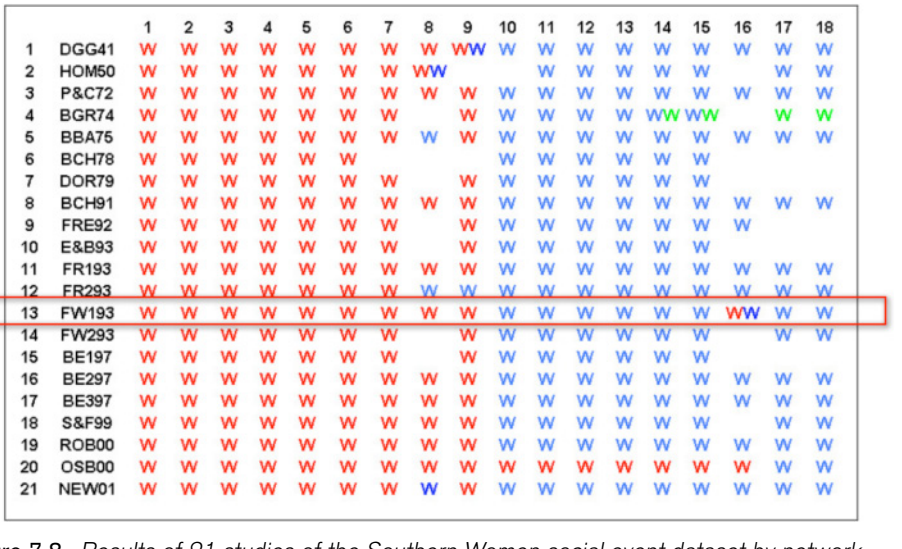

Figure 7-8. *Results of 21 studies of the Southern Women social event dataset by network scientists (Freeman 2003)*

Look at the various membership groupings in Table 7-1. Most of the studies came to highly similar conclusions, and all found two distinct clusters in the data. However, there is not total agreement about who is in each cluster, especially for women #8–18.

This table illustrates membership groupings well, but it does not reveal network roles and social distances. The network map in Figure 7-7 does reveal the nuances of the social structure and shows the *points of failure* in the network—that is, where it is most likely to break down. For instance, if woman #3 were to move away, the network would be disrupted the most. It would be interesting to see how both woman #4 and woman #9 would respond to the exit of woman #3.

Social Graphs of Amazon Book Purchasing Data

Amazon.com allows easy access to summary purchase data (transaction data is aggregated to prevent individual identification). The book purchasing data Amazon provides forms a similar network dataset to the event network in Figure 7-3. Instead of attending the same social events, on Amazon, people are connected to one another by purchasing the same books. In both cases, connections are made because certain people make the same choices as others.

On each item's page, Amazon provides the following information:

> Customers Who Bought This Item Also Bought

When people buy two items, an association is formed between those items. The more people purchase both items, the stronger the association is and the higher on the list the *also bought* item appears. Although usually people are represented by nodes, in this case Amazon's customers are the links in the network, and the items they purchase are the nodes. Consequently, Amazon is able to generate a network that provides significant information about its customers' choices and preferences, without revealing any personal data about the individual customers. Patterns are revealed, while privacy is maintained. With a little data mining and some data visualization, we can get great insights into the habits and choices of Amazon's customers—that is, we can come to understand groups of people without knowing about their individual choices.

Determining the Network Around a Particular Book

One of the cardinal rules of human networks is "birds of a feather flock together." Friends of friends become friends, and coworkers of coworkers become colleagues. Dense clusters of connections emerge throughout the social space. In the social networks we visualize, we see those birds of a feather near each other on the map.

Let's take a look at a popular computer book available via Amazon: Toby Segaran and Jeff Hammerbacher's *Beautiful Data* (O'Reilly). Among other information, the book's Amazon page provides a product description, publication details, and a brief list of "also bought" books. What does this list tell us about the book we are viewing? Being a student of networks, my inquiry about this book did not stop at the *also bought* books listed on this web page (one step in the network). I wanted to know what would happen if I followed the links to each of those books and joined the lists I found there into a network (one and two steps in the network).

Key to understanding the dynamics of networks is the ability to perceive the *emergent patterns of connections* that surround an individual node, or that are present within and around a community of interest. I wanted to see the network in which my book of interest was embedded. Seeing those connections can provide insight into the *network* neighborhood—the network surrounding this book—which can help a consumer make a smarter purchase.

Tracing the network out two steps from the focus node is a common procedure in social network analysis when studying *ego networks*. An ego network allows us to see *who* is in one's network neighborhood, *how* they are interconnected, and *how* this structure may influence the ego—the focus node.

As I collected the *also bought* books around *Beautiful Data*, I wondered:

- What themes would I see in the books and in their connections?
- What other topics interest the readers of *Beautiful Data*?
- Will *Beautiful Data* end up in the center of one large, massively interconnected cluster or be a part of one distinct community of interest amongst several?

Figure 7-9 shows the book network surrounding *Beautiful Data*. Each node represents a book purchased on Amazon. A gray line links books that were purchased together, with the arrowhead pointing in the direction of the *also bought* book. The red nodes represent other books published by O'Reilly Media, while the yellow nodes represent books from other publishers.

In networks, it is not the number of connections one has, but where those connections lead, that creates advantage. The golden rule in networks is the same as in real estate: *location, location, location*. In real estate, what matters is physical location: geography. In networks, it is virtual location, determined by the pattern of connections surrounding a node.

The nodes in Figure 7-9 self-organize, in the graph space, by their ties to *also bought* books. This allows similar books to self-organize together to form clusters of like topics, which reveal the human communities of interest behind the book clusters. In Figure 7-9, two obvious groupings cling together by topic:

- The bottom-right grouping is all about programmers and programming.
- The grouping at the top of the graph is all about the Semantic Web.

Although clusters emerge in Figure 7-9, they are not as obvious as some others that we will see later; these clusters are intermixed and overlap, especially around other books about modern programming methods and processes.

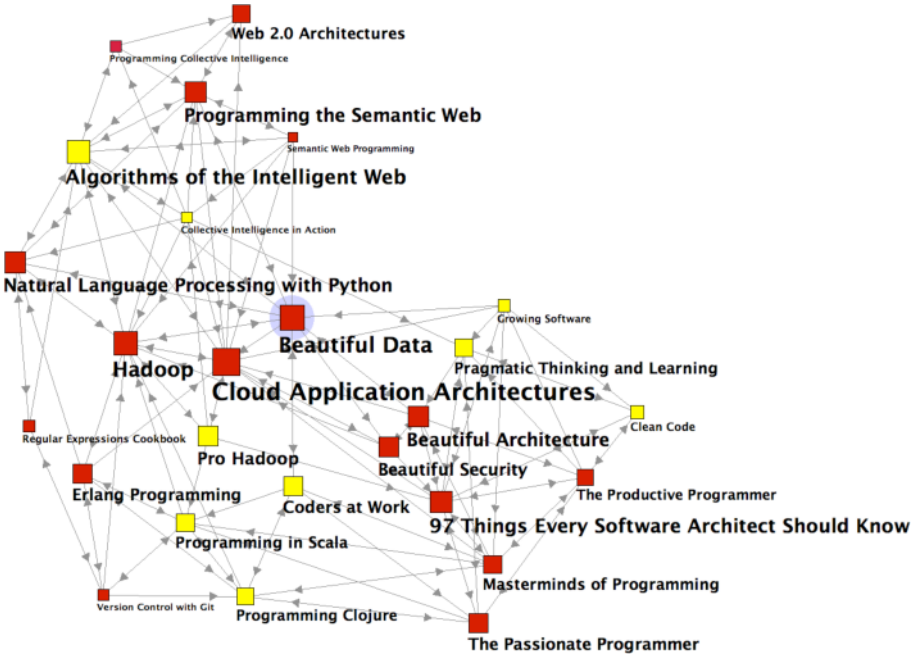

Figure 7-9. *The network neighborhood of books surrounding Beautiful Data*

In addition to clusters of like topics, in Figure 7-9 there are clusters around the publishers, designated by the node colors: red books connect to other red books and yellows connect to other yellows. This indicates that people who like O'Reilly books tend to buy other O'Reilly books. Node size also appears to form a weak pattern of connectivity across similarly sized nodes. Large nodes, the nonlocal influence across the graph, connect to other large nodes, while medium and small nodes often connect to one another. This is a pattern we often see in human networks—again, birds of a feather flock together. It is not a pattern we see with the physical structure of the Internet, though, where many small nodes connect to a few very large nodes, creating an obvious hub-and-spoke pattern. That is often referred to as a *scale-free network*.

Next, I examined the network measures of each node/book, to see which nodes were well positioned in the web of connections. Since this is a directed network, much like the World Wide Web, I calculated influence metrics similar to Google's PageRank. These metrics were calculated using both direct and indirect links around each node. Like on the Web, a better-connected node transfers more influence. These metrics do not reflect sales volumes or the popularity that quantity conveys; rather, they reveal what thousands of Amazon purchasers feel belong together—what the "birds of a feather" books are. The larger nodes have greater influence in this community of interest based on the pattern of *also bought* purchases.

Another common network measure is *structural equivalence*. This measure reveals which nodes play a similar role in a network. Equivalent nodes may be substitutable for one another in the network. As an author, I would *not* like my book to be substitutable with many other books! As a reader, however, I would like equivalent choices. In Figure 7-9, the two books with the most similar link pattern to *Beautiful Data* are *Cloud Application Architectures* and *Programming the Semantic Web*.

Another value-added service that Amazon provides is reader-submitted book reviews. A person considering the purchase of a particular book may be aided by the many reviews that accumulate. Unfortunately, the reviews can be skewed: an author with a large personal network can quickly get a dozen or more glowing reviews of his latest book posted to Amazon, and a reader with a grudge can do the opposite. Doing comparison shopping based on reader reviews alone may, therefore, be misleading.

The book network map may be a better indicator than individual reviews of which other books to buy. Books linked from many other similar books reflect critical choices made by purchasers, who spent money on those books. Surely this behavior is not random; it is executed on the basis of thought and comparisons. A purchase decision is the best review of all, even if it is never written.

The book network maps I've shown are designed to eliminate the *peripheral nodes* in the network (i.e., those with very few connections). The network map in Figure 7-9 shows a *3-core network*—a network in which each node has a minimum of three connections to other nodes. To achieve this, all nodes with only one or two incoming links were removed. These were nodes that led to other communities of interest, that represented new or very old books, or that had very few *also bought* links from this community.

Putting the Results to Work

These community-of-interest maps can also work in a similar capacity with other consumer items. If I am not familiar with a product, an author, an artist, a vintage, a brand, a movie, or a song, I would like to be able to judge it by the company it keeps—its network neighborhood. Here are the relevant questions to ask:

- What nodes point to this item?
- What communities is it a member of?
- Is it central in the community?
- Does it bridge communities?
- Are there equivalent alternatives?

It appears that as a customer of Amazon, I can make smarter decisions by viewing the *embeddedness*—the context within the network—of various items Amazon sells in different communities of interest. Other vendors, such as Netflix and Apple's iTunes,

probably do similar analysis before recommending a movie or a new song or artist. By gathering information on thousands of customers and what they choose and organize together, a vendor can form a product-to-product network like that in Figure 7-9, or even a person-to-person network like that in Figure 7-7. Both maps will indicate likely influence patterns and what it makes sense for customers to purchase/rent/download together.

Here are some network rules of thumb that we can distill from the Amazon analysis:

- If you have read one nonfiction book of a structurally equivalent pair, you may not be in a rush to read the second, since the second book probably covers the same information as the first book. On the other hand, you may wish to read a large number of structurally equivalent fiction titles (can't get enough of those cyber-thrillers!).
- If you liked books A, B, and C and want to read something similar, find which books are linked to A and B as well as C. You can only see this in the network diagram; you cannot see these linkages in Amazon's individual lists unless you open three browser windows and compare the lists yourself.
- If you want to read just one book about topic X, find the book with the highest network influence score in the cluster of topic-X books. This follows the Google PageRank approach and may reveal a book with excellent "word of mouth" appeal.
- If the book you are looking for is not in stock, find which other books are structurally equivalent to that book. These will provide similar content and may be available.

A book author and/or publicist could use her knowledge of existing book networks to position a book where there is a *hole*, or gap in the network. A publisher could review evolving book networks, which may change weekly, to adapt its marketing efforts. Amazon, of course, is still the big winner: it has all the data, and a rich upside of hitherto untapped possibilities for analyzing the data and applying the findings.

Social Networks of Political Books

Visualizing book networks on Amazon not only helps us choose which books to purchase, but also gives us insights into larger trends and patterns in a particular sphere of interest. One area that is ripe for exploration is politics. Purchase patterns on Amazon often reflect the results of countrywide surveys of political beliefs and choices.

Two books are connected in the book network if Amazon reports that they were frequently bought together by the same consumer. I don't arrange or color the nodes before feeding the *also bought* data through my social network analysis software,

InFlow 3.1.* The software has an algorithm that arranges the layout of the nodes based on each node's connections. Once the software finds the emergent pattern and identifies any clusters, I review the books in each cluster and then see whether they naturally cluster as blue, red, or purple (my coloring scheme follows the conservative-as-red and liberal-as-blue convention that became popular in the United States during the 2000 presidential election; purple is a combination of red and blue and is used to describe books that fall between the two popular political camps).

I have been doing a social network analysis of the purchase patterns of political books since 2003. Unsurprisingly, from my very first mapping I saw two distinct political clusters: a red one designating those who read right-leaning books and a blue one designating those who read left-leaning books. In my 2003 network analysis, I saw just one book holding the red and blue clusters together. Ironically, that book was named *What Went Wrong*. This map is shown in Figure 7-10.

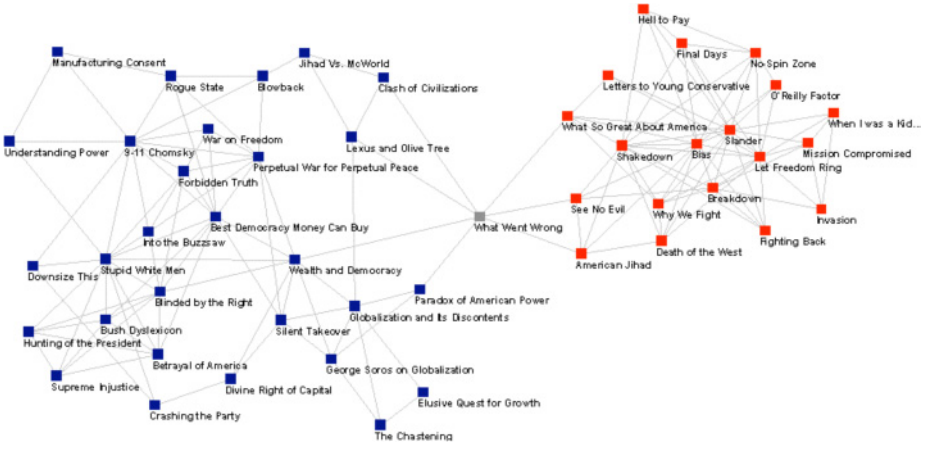

Figure 7-10. *Divide of political books in 2003*

In the 2004 map (Figure 7-11), constructed several months before the 2004 U.S. presidential election, several books held the two clusters together. Again, at least with the better-selling books, there was very little crossover between the right and left camps: people on each side appeared to be reading more and more books that supported their existing frames of mind. This is not to say that no readers were reading both red and blue books, but they appeared to be in the minority. I looked only at Amazon's best-selling books and at the most common *also boughts* for each book, focusing on the most frequent and intense interactions (as when examining the strong ties in a human network). A deeper look into the Amazon data (if Amazon permitted it) might reveal

* See *http://orgnet.com/inflow3.html*.

these weaker, less frequent, connections amongst blue and red books. I would expect to see a small minority reading books on both sides—many might be in academia, teaching or taking courses where both sides of an issue are presented and debated.

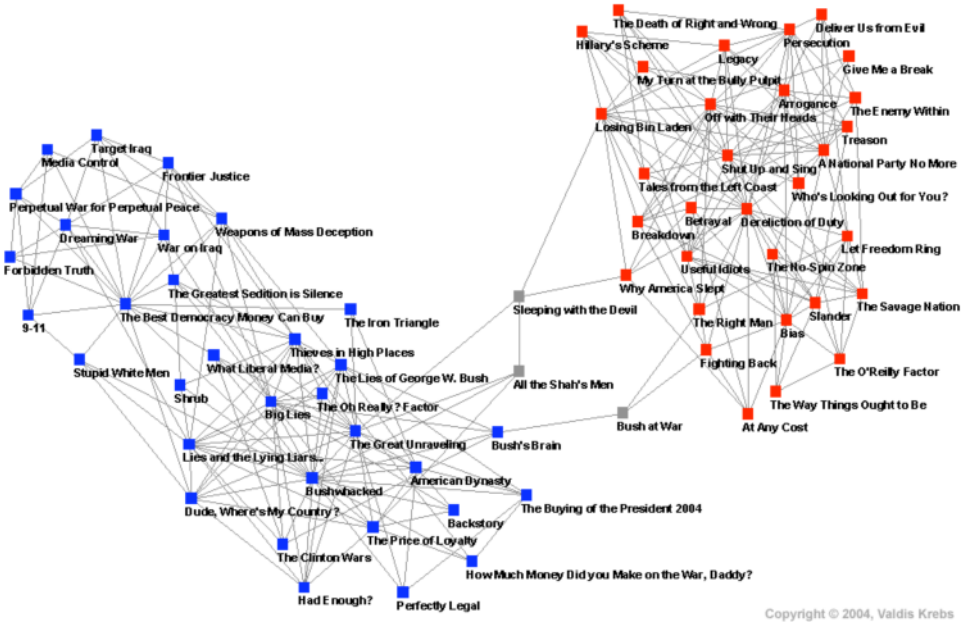

Figure 7-11. *Divide of political books in 2004*

I continued to create these political book maps using Amazon data from 2005 through 2007, and I kept getting the same strong red/blue divide. *The books changed over time, but the overall network pattern remained the same*. How strong was this pattern? To test it, I experimented with my data collection approaches—were the strong patterns an artifact of my methods? No! Regardless of the data collection method, as long as I followed accepted practices—such as "snowball sampling" (Heckathorn 1997)—the results showed strong red and blue clustering. Occasionally a different collection method would result in a few new books sneaking into the mix, but the overall pattern remained stable. The emergent political book network pattern was *not* sensitive to data collection methods and cutoffs, indicating that the pattern was strong and persistent.

In 2008, with the U.S. presidential election approaching, I decided to take several snapshots of the political network. How would it change as we moved closer to Election Day? I captured the network at three critical junctures:

- At the end of the primary season
- After the last convention
- Right before Election Day in November

I expected the red/blue divide to persist, but wondered if any interesting patterns would appear as the presidential election process moved through its phases.

In June 2008, after the major party candidates had been chosen via the primary process, I turned again to the predictive patterns of partisan political polemics. At the Iowa caucus in January of that year, Obama had said, "we are not a collection of red states and blue states, we are the United States of America," and McCain proclaimed his purple "maverick" roots. But what did the book data tell us?

Figure 7-12 was created during June 2008. As a little experiment, I added a new color: light blue. According to the Amazon sales data, these books cluster with the other blues. But looking at the titles and authors, they do not fit in with the common blue themes and the supporters of previous iterations of blue nodes. At this point in time, popular conservatives, independents, and libertarians were all finding more connection with the blue readers than with the red readers. The reds had only George Will bridging them to the rest of the U.S. political world, and a split on the right between the "old conservatives" and the "neo-cons" emerged, with the old conservatives more aligned with the progressives than with the neo-cons in the summer of 2008.

Figure 7-12. *Political book purchase patterns during June 2008*

In August 2008, several anti-Obama books appeared. A new pro-Obama book, with a foreword written by Obama himself, was also in prerelease and being sold on Amazon. Figure 7-13 reveals who was reading these books. The pro-Obama book, *Change We Can Believe In*, is solidly in the blue cluster, indicating that people who had already

purchased pro-Obama books were also purchasing this positive book. Similarly, the anti-Obama books—*The Obama Nation* and *The Case Against Barack Obama*—were primarily being purchased by people who had already purchased other anti-Obama books. One of the anti-Obama books, though, is connected to one of the purple books, *The Late Great USA*. Could some undecided voters, not happy with the state of the country, have been reading this book to make up their minds about Obama?

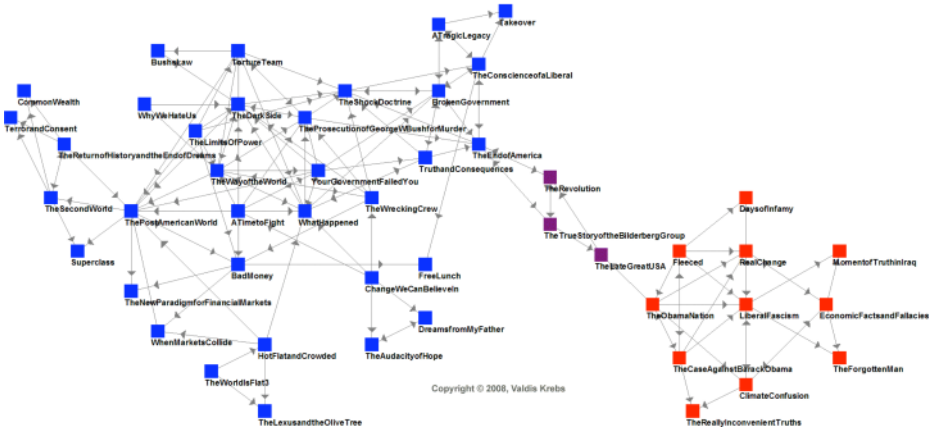

Figure 7-13. *Political book purchase patterns during August 2008*

No books on McCain, either pro or con, were amongst Amazon's best-selling political polemics. Did people already know enough about him at this point in the election cycle, or were they not interested in him? The pattern of connections between the books in the map in Figure 7-13 indicate that the most influential political books at the end of the summer of 2008 were *What Happened* and *The Post American World*—neither addressed the current election! *What Happened* was written by the former press secretary for George W. Bush, but it was being purchased by the blue readers only.

Social network analysis and data mining/visualization provide us with two categories of outcomes:

- Expected versus unexpected results and insights
- Positive versus negative results and insights

These categories intersect, as illustrated in Figure 7-14. In the hundreds of social network analysis projects I have participated in, I have found that clients typically most enjoy seeing what they did not anticipate—the unexpected (and especially *negative unexpected*) patterns that can lead to problems.

Positive Expected	Positive Unexpected
Negative Expected	Negative Unexpected

Figure 7-14. *Discovery matrix for social network analysis*

Let's examine our last graph using the discovery matrix in Figure 7-14. In late October 2008, as both presidential campaigns sprinted toward the finish line, I took one more look at the political books being purchased and the patterns they created. The pre-election network map is shown in Figure 7-15. A few unexpected patterns emerge in this map, along with one expected pattern.

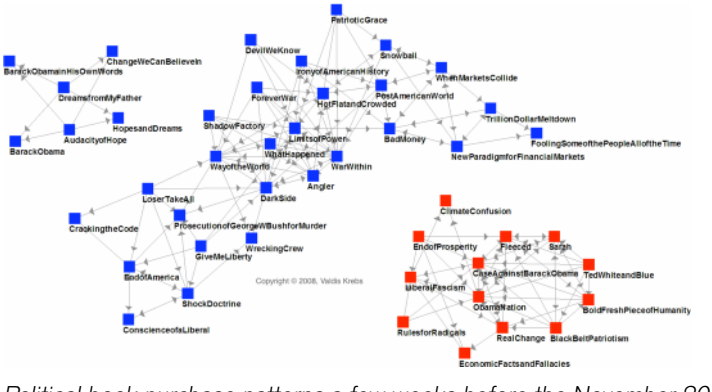

Figure 7-15. *Political book purchase patterns a few weeks before the November 2008 election*

Unlike in all the previous maps, there are *no* bridging books between the red and blue clusters—the two sides are totally separate! Red and blue have nothing in common! This pattern reflects the immense polarization and animosity evidenced in the campaign rallies in the run up to the election. Political issues and the great economic problems of the time were not being discussed. This pattern can be classified as a *negative expected* based on the daily actions of each campaign.

Another revelation of the visualization in Figure 7-15 was that right-leaning readers had been buying the key book of community organizers, *Rules for Radicals*. This same group had mocked community organizing! Why were right-of-center readers buying this book, which was normally popular with a left-of-center audience? Was the right trying to figure out why Obama's campaign, based on community organizing principles, had been so successful? This was an *unexpected* pattern, but whether you think it should be classed as positive or negative probably depends on which side you are on.

A final unexpected pattern was that those buying positive books about Obama were not buying other political books. The "about Obama" cluster is disconnected from the other clusters that contain political polemics. This pattern may indicate that these readers are interested only in Obama and this election, not in politics in general.

An *expected* pattern also jumps out from this pre-election political book network map. Since 2004, there have been more registered Democrats than Republicans, so it makes intuitive sense that there are more blue books. In contrast, the right focuses on fewer books to get its message across (the book network map does *not* reflect volume of books sold, so it is possible that readers on the right actually buy a greater volume of fewer books—we don't know, as Amazon does not reveal this data). This is probably viewed as a *positive expected* pattern by both sides, but for different reasons. The right is likely to view its approach as more focused, while the left interprets it as the opposition lacking a variety of opinions. Conversely, the left is likely to view the larger number of books on its side positively, as representing a diversity of opinions, while the right may view it as indicating a scattered and unfocused message.

Conclusion

As the visualizations presented in this chapter have illustrated, *our choices reveal who we are, and whom we are like*. The decisions we make identify not only certain aspects of ourselves, but also what groups we belong to. Since "birds of a feather flock together," our choices provide many insights into the behaviors of others in our groups. In the future (on the Web, for example), many of our choices may not be conscious: our smartphones will communicate with other nearby smart devices looking for ways to connect with their owners. These devices may be programmed to look for the patterns we have examined here. A few brave souls may program their devices to selectively break the typical patterns in which they are embedded—for instance, a red-book reader could strike up a conversation with a blue-book reader after their devices reveal the opportunity to exchange viewpoints.

The Amazon data illustrated that we can gain deep insights into the political choices and behaviors of different groups without knowing anything about the individuals belonging to those groups. Private data does not need to be revealed for us to understand large-scale political patterns based on book purchases. Even more amazing, this data, along with the simple visualizations created to display it, matched the findings of expensive nationwide surveys of potential voters. An hour collecting and mapping

Amazon data gave some of the same insights as thousands of hours spent collecting and analyzing voter survey and interview data. The Pareto 80/20 rule works well here: we get 80% of the insight for much less than 20% of the time invested—an excellent payoff when properly matching data mining with data visualization!

References

Davis, Allison, B.B. Gardner, and M.R. Gardner. 1941. *Deep South: An Anthropological Study of Caste and Class*. Chicago: University of Chicago Press.

Freeman, Linton C. 1979. "Centrality in social networks: I. Conceptual clarification." *Social Networks* 1: 215–239. *http://moreno.ss.uci.edu/27.pdf*.

Freeman, Linton C. 2003. "Finding social groups: A meta-analysis of the southern women data." In *Dynamic Social Network Modeling and Analysis*, eds. Ronald Breiger, Kathleen Carley, and Philippa Pattison. Washington, DC: The National Academies Press. *http://moreno.ss.uci.edu/85.pdf*.

Freeman, Linton C. 2004. *The Development of Social Network Analysis: A Study in the Sociology of Science.* Vancouver, Canada: Empirical Press. *http://aris.ss.uci.edu/~lin/book.pdf*.

Heckathorn, D.D. 1997. "Respondent-driven sampling: A new approach to the study of hidden populations." *Social Problems* 44: 174–199.

Mayo, Elton. 1933. *The Human Problems of an Industrial Civilization*. New York: MacMillan.

Moreno, Jacob L. 1934. *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Foreword by Dr. W.A. White. Washington, DC: Nervous and Mental Disease Publishing Company.