Paper Summary: Newman, The Structure of Scientific Collaboration Networks

January 27, 2004

M. E. J. Newman, "The structure of scientific collaboration networks," Proceedings of the National Academy of Sciences USA 98(2):404-409, January 16, 2001.

The contribution of this work is an analysis of the network structure of several different real-world scientific collaboration communities. Four databases, representing large numbers of published papers in biomedical, theoretical physics, high-energy physics, and computer science research, are studied.

Previous work cited in this area includes Milgram's somewhat informal "six degrees of separation" social-network study, studies of relatively small social networks based on subjective survey data, and studies of large physical (but non-social) networks. There has also been some work on co-authorship and co-citation, but not using as extensive a citation database as this work.

Oddly, there are no article titles in the reference list. The references I'd look up would be Milgram's 1967 "six degrees" study (obviously a seminal result); the Watts and Strogatz 1998 Nature paper (apparently a key result, published in a major journal); and the Watts *Small Worlds* book (recent enough to likely be a good survey/introduction to the field).

The key findings are:

- The average degree distribution does *not* follow a power law distribution, but instead exhibits a "power law with exponential cutoff." Interestingly, *all* of the communities exhibit the same distribution, but with different τ (exponent) and z_c (cutoff) parameters. **Question:**The authors speculate that the cutoff is due to the finite time ranges covered by the databases, but I didn't entirely understand this discussion. **Question:** Is the P value just the confidence level for statistical significance? (Not the same probability value in Equation 1? That's a bit confusing...) Is R the correlation coefficient?

- "Six degrees of separation" seems to apply, with an average minimum path length between a random pair of scientists of approximately 6. Interestingly, computer scientists are less connected (average distance is 10). The diameter (maximum path length) of all of the networks is roughly 20. **Question:**I didn't follow the argument about the "correlation between the measured distances and the expected log N behavior."

- The communities are fairly well connected, with typically 80 or 90% of authors falling into a single connected component (the "giant component"). Newman found that the second-largest connected component is much smaller, and concludes from this that scientific collaboration networks are well connected. **Question/comment:** He draws this conclusion because networks that *are* highly connected exhibit this property (of one very large component), but it seems like a rather indirect argument. The clustering would seem to be a better direct indicator of "tight connectedness."

- Most of the communities exhibit excess clustering (i.e., two authors who have collaborated with a third author are significantly more likely to have collaborated with each other) – but this effect is not seen in the biomedical community.

The work seems to be solid, but not especially innovative. The metrics used are all from the existing literature; they have been applied to new datasets. Some of the conclusions and speculations are interesting, and might point the way to further investigations.

The paper is very well organized and well written.

Parberry Classification: Tinkering?

Parberry Analysis:

- **Correctness:** Basically seems correct. A few of the conclusions seem like a stretch, but most such conclusions are marked as speculations.
- **Significance:** Moderate to low. The "exponential with cutoff" finding with regard to degree distribution seems interesting; the rest doesn't seem to be very important.
- **Innovation:** Fairly low; mostly just applies existing analysis methods to new data.
- **Interest:** Moderate to high. Scientific collaboration networks are a "hot item" in CS/AI/IR these days, so the analysis and conclusions may be of interest, even though they're not particularly innovative or conceptually significant.
- **Timeliness:** High.
- **Succinctness:** High. The paper is very concise and yet understandable.
- **Accessibility:** Moderate to high. I had a bit of trouble following some of the analysis on first reading, but generally speaking it's very accessible.
- **Elegance:** Moderate to high. No proofs, but the analysis is reasonably elegant.
- **Readability:** Moderate to high. A few technical details are glossed over (presumably under the assumption that the reader is already familiar with power laws and the like).
- **Style:** High. Very well organized, well written and readable.
- **Polish:** High. Very clean.