Effect of Distance on Collaboration in the Internet Age Is not Scale Free: Distance Matters *and* Distance Doesn't Matter

Henry Kim

Schulich School of Business, York University, Toronto, Canada.
hmkim@yorku.ca

Juyong Park

Department of Physics, University of Michigan, Ann Arbor, MI, USA
juyongp@umich.edu

Eleanor Wynn

Intel Corporation, Portland, OR USA
eleanor.wynn@intel.com

Introduction

Has the Internet truly managed to bridge and mitigate the effect of geographical distance in enabling collaborative work? Most would say 'yes', but it is important to go beyond conjectures and anecdotal evidence and examine objective data on how it has brought people together, or perhaps not, for knowledge-intensive, decentralized collaborative work. A prime example of this kind of work is collaborative research. Research can be done between those co-located within a lab, but very often is done between researchers in different labs or departments, different institutions (or universities) or different countries. Research collaborations, between those at universities in particular, are initiated in a self-organizing manner without a central dictate. Especially in these situations, distance may have a positive or negative effect on whether collaboration occurs.

One proposed perspective is that the Internet would enable the lowering of the cost of collaboration so much that researchers would form true virtual communities that supplant their geographical work communities in importance. Science, then, would be balkanized along narrowly-defined and insular interest clusters that are not mitigated by co-location [VanAlstyne and Brynjolfsson, 1996]. Information technologies cannot usually compensate for the rich, synchronous communication needed for distant collaboration except when collaborators have common ground, do loosely coupled work, have enabling technologies, and are ready to use them [Olson and Olson, 2000]. Much of academic research may fall in this category, particularly research not based in the physical sciences.

However, geographical distance has traditionally been an impediment to academic collaboration; studies show that occurrences of co-authorships decrease exponentially as distance between collaborators increases [Katz, 1994, Nagpaul, 2003]. A survey of economics research show that this is still the case, but also shows that researchers with similar interests are now more likely to overcome distance barriers and collaborate than in the 1970's and 80's [Rosenblat and Mobius, 2004]. This is consistent with the theory of the balkanization of science. So has the effect of distance in academic research been largely mitigated, then? Have the Internet and the Web fulfilled their promise to

academics, who were the first to use and study the technologies, and who best fit the profile of those for whom distance should *not* matter, or matter the least?

In this paper, we analyze the collective collaboration behavior of the very researchers who were the first to study how organizations used information technologies: the Management Information Systems (MIS) research community. These researchers are most often found in business schools and their works are published in venues like the *Communications of the ACM*. We collected bibliographic data from the two premier journals of this community—*MIS Quarterly* and *Information Systems Research*[1]—from 1985 to 2001. The dataset comprised of 679 articles written by 873 different authors who were affiliated with 364 distinct institutions, the great majority of which were universities. There were on average 2.17 authors and 1.78 affiliations per article. We augmented this dataset by calculating the pair-wise geographical distance between affiliated institutions associated with a paper[2]. So for one affiliation, there were no inter-institutional collaborations; for two institutions, one distance was calculated; for three institutions, we calculated three distances; for four institutions, we calculated six distances, and so on. Though we would have preferred to calculate distances between authors, nearly all large scale data sources including ours (Web of Science™) do not facilitate capturing data that uniquely identify the author of an article with his/her exact affiliation(s).

---

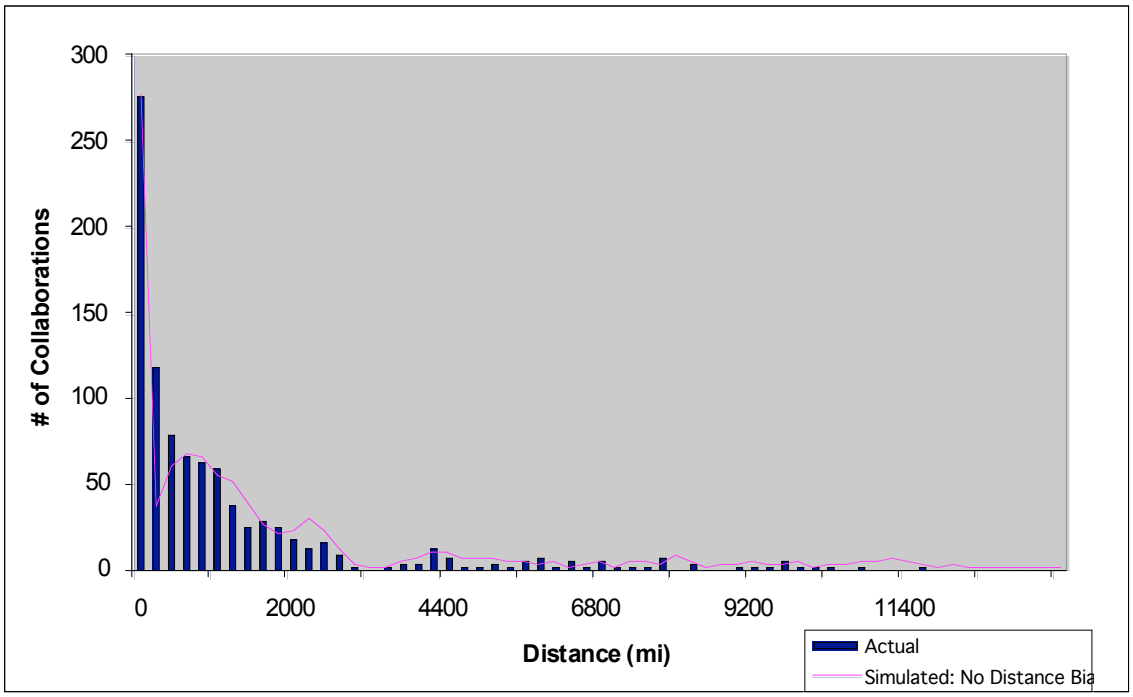
[1] *Information Systems Research* was first published in 1990.

[2] The calculations were done by referencing the latitude and longitude of the cities of the institutions and using a standard geodesic calculation formula to approximate the distance between the institutions.

Data Analysis

The chart below shows that our results were consistent with others': The number of collaborations decreased exponentially as geographical separation (distance) increased. We also ran a simulation wherein the collaborative ties were 'shuffled' so that each institution still had the same number of collaborations as before, but the collaboration partners were now random. For example, if USC, UCLA, and Michigan published and collaborated with other institutions with the same likelihood, then USC was as likely to publish with Michigan as it was with UCLA (i.e. it was blind to others' distances from itself). The distribution that the simulation generates reflects that in the US—where the bulk of the schools are—there are more highly publishing institutions that a given institution has collaborated with within a distance of approximately 800-1000 miles, beyond which there aren't as many such highly publishing schools. Though the parts on the left differ for the actual and simulated distributions, they both have fast-declining tails; collaborations at over 5000 miles, for instance, are rare (accounting for around 0.5% of all inter-institutional collaborations). Comparing the actual versus simulated distributions, the effect of distance appears very compelling and does not appear to be mitigated at all at the range of very large distances.

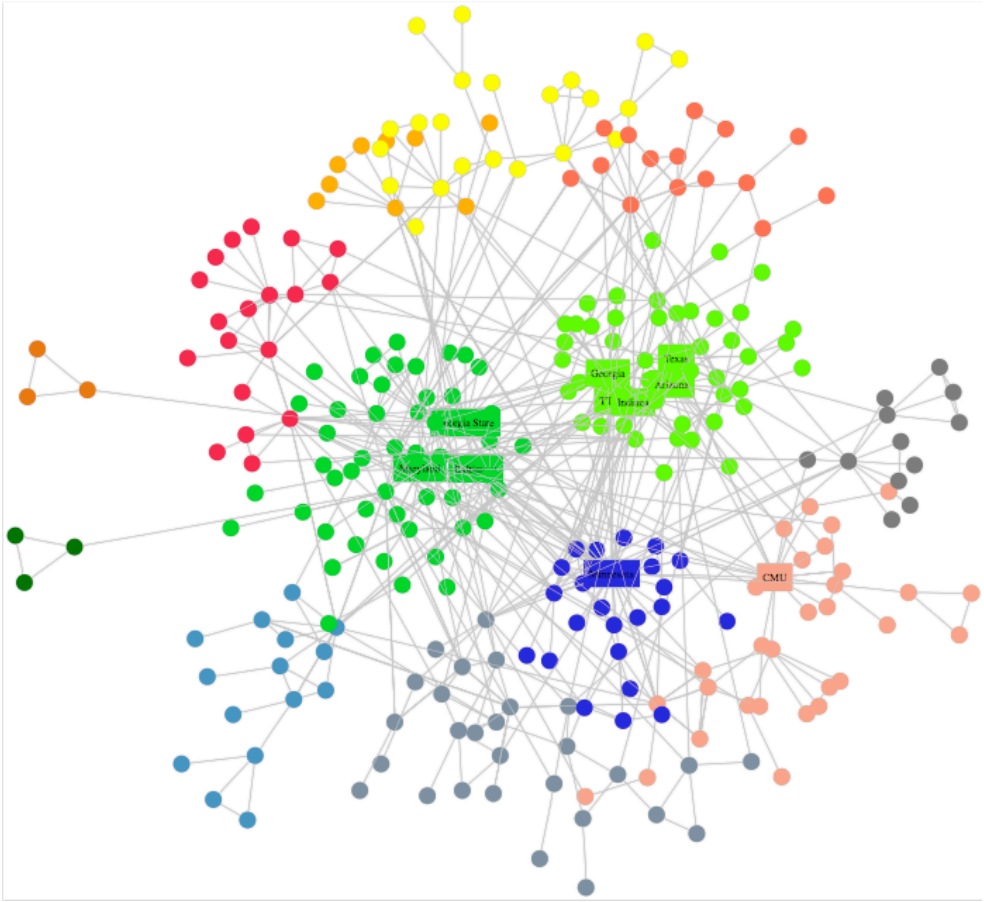
Figure 1: Effect of Distance on Inter-Institutional Collaboration[3]



To better understand the nature of inter-institutional collaboration in this field, we grouped universities according to their propensity to collaborate using a social network algorithm used in determining community structures proposed by Newman [Newman, 2004]. Based on the intuition that a set of nodes that belong to a common group will have a higher number of connections among themselves than to nodes in different groups, the algorithm maximizes a quantity called "modularity" that reflects the ratio of the intra-group connections to inter-group connections. This algorithm has been effectively applied to identify community structures, for example, within the collaboration patterns of physicists. The figure below shows the communities that result from executing the algorithm on the dataset.

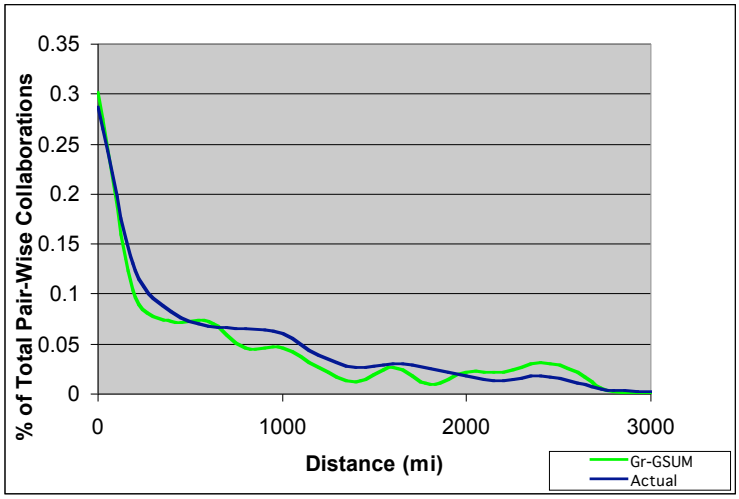
[3] The distances are classified into 200 mile categories; i.e. 0-200 miles, then 201-400 miles, etc.

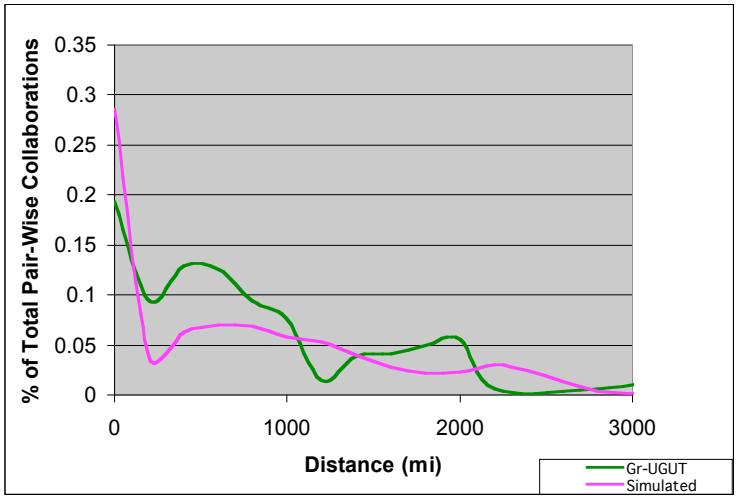
Figure 2: Communities of MIS Research Universities



The schools that are mentioned are those that are the most published in this field. We performed further analysis on two of the most prolific communities—the group identified as Gr-GSUM comprised of 52 institutions including Georgia State and Maryland; and Gr-UGUT comprised of 49 institutions including Georgia and Texas at Austin. We charted the collaboration distances of the schools in each group. Interestingly, Gr-GSUM exhibited the exponential decay pattern à la the distribution of the whole dataset, while Gr-UGUT exhibited the distance-unbiased pattern à la the simulated distribution.

Figure 3: Effects of Distance on Inter-Institutional Collaboration by Communities





The field of MIS emerged from the University of Minnesota, and so it is viewed as the true bridge institution in this field and at the same time the most prolific. Not surprisingly then, given the community structuring algorithm we used, it was not part of our two study groups but rather was the locus for another group. The next 11 out of the top 17 most productive schools were included in Gr-GSUM (UC-Irvine, Georgia State, Maryland, British Columbia, Harvard, Pittsburgh, and Florida International) or Gr-UGUT (Georgia, UT-Austin, Arizona, and Indiana). A researcher from these 11 schools (out of 364 institutions) was a co-author on 27% of all papers. An author from one of the six top

schools in the first group was involved in 73% (132 out of 180) of the group's articles; it was 57% (77 out of 134) for the four top schools in the second group.

When we created weighted maps of schools in these groups—with regions with more productive schools displayed with more emphasis—a distinguishable pattern emerged. The maps highlight that except for the Georgia schools, Gr-GSUM contained productive schools near the East and West coasts of North America while Gr-UGUT was comprised of more inland schools. To help read the maps, we present these examples: the blue bar represents the region around 45°N-120°W where University of British Columbia is located and the orange bar represents the region around 34°N-100°W where UT-Austin is. The bars represent the number of articles published by schools in a latitude-longitude section.

Figure 4: Weighted Geographic Distribution of North American Schools in Gr-GSUM

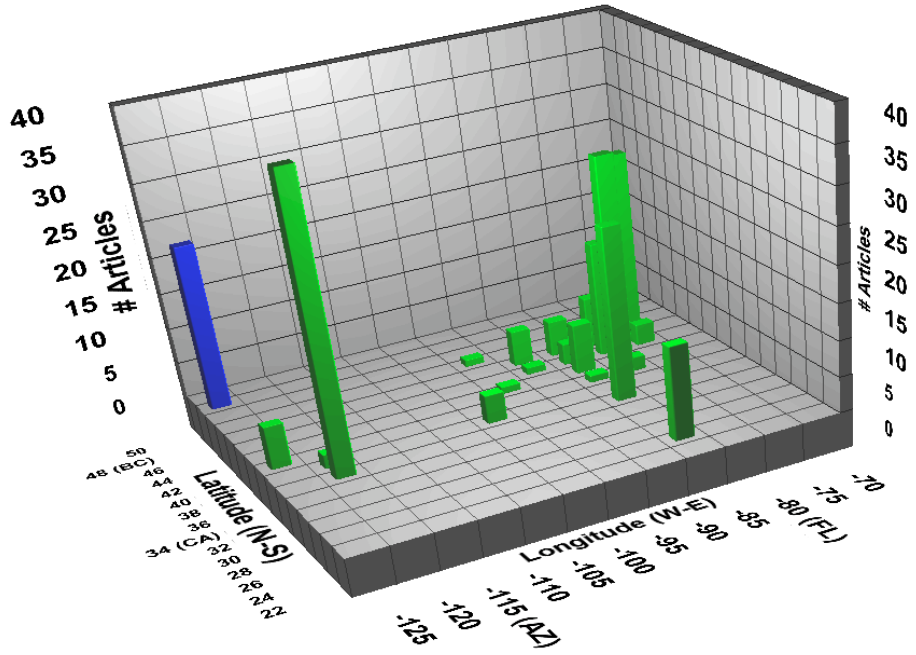
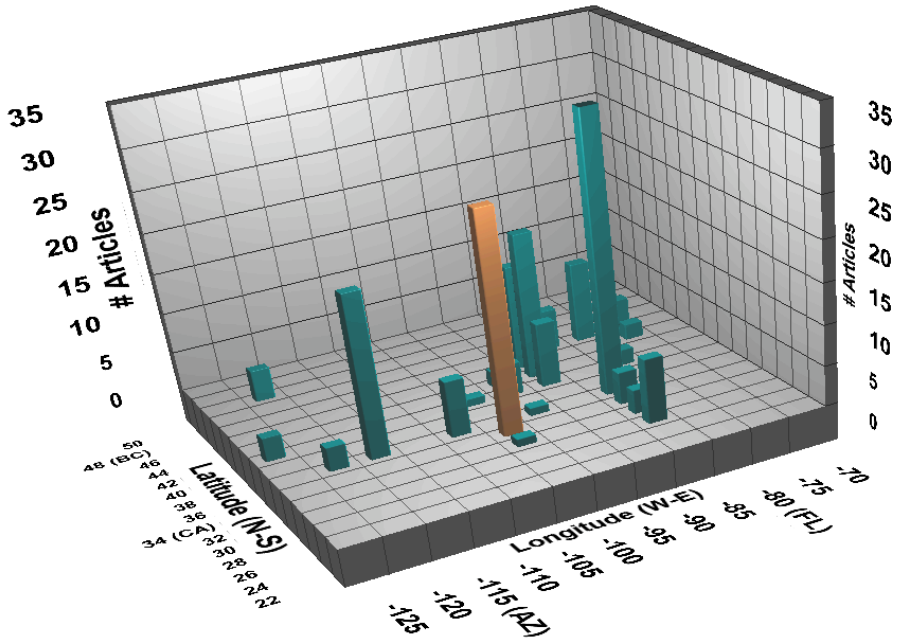


Figure 5: Weighted Geographic Distribution of North American Schools in Gr-UGUT



The association between geographical and collaboration distance distributions now seems more clear. When an author can collaborate with a colleague or others in many schools nearby—as would be the case in the Eastern US states or California—then that author is less inclined to engage in longer distance collaborations. This may explain the exponential decay in the collaboration distance distribution for Gr-GSUM. However, when authors are located in places like the Western US states where the population density of potential collaborators is lower, they may be more willing to engage in longer distance collaborations to overcome their geographical disadvantage. This result is very interesting because a simple macroscopic factor, geography, seems to express much of the rationale for the community structuring even though we did not factor in more complicated microscopic factors related to the article (e.g. general topic), individual (e.g. where a researcher has schooled and worked), the journal (e.g. who the editors are), or institution (e.g. specific competencies within the MIS field in which the school specializes).

However, this simple explanation does not adequately explain why University of Georgia and Georgia State, which are less than 100 miles apart and in the top 5 in terms of productivity, belong to different groups. Below, the North American regions of the schools that collaborate with the two universities are mapped. The maps show that Georgia collaborates with East and West Coast schools in a similar way to Georgia State, but additionally collaborates with schools in the West and Midwest (shown in yellow). Thus Georgia is grouped in Gr-UGUT.

Figure 6: Geographical Distribution of Institutions that Collaborate with Georgia State

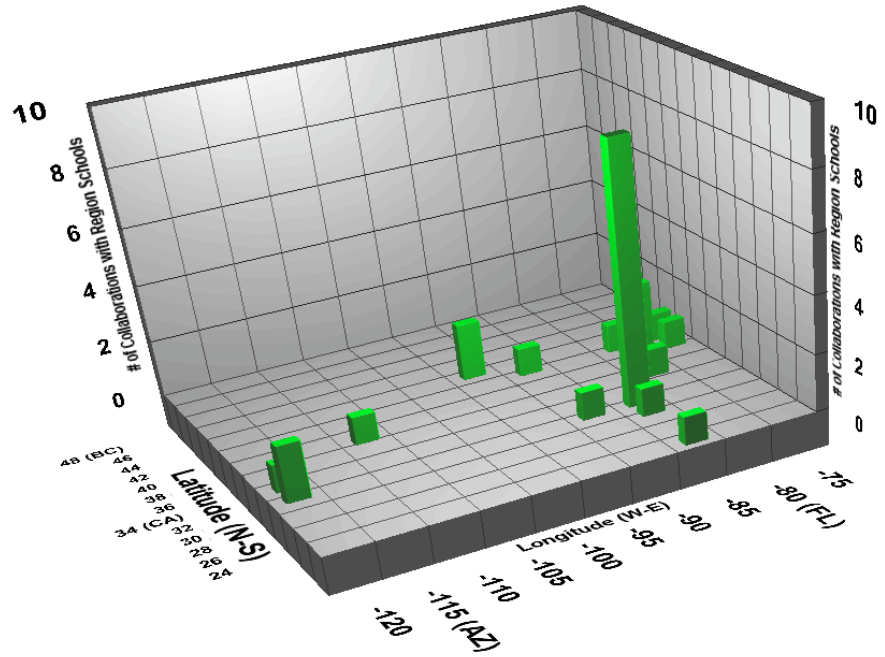
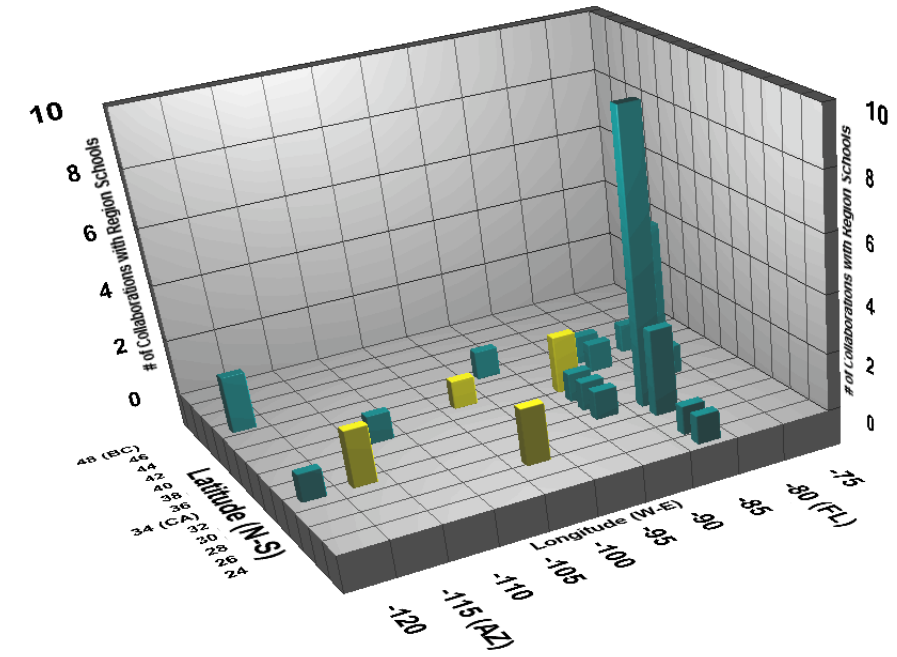


Figure 7: Geographical Distribution of Institutions that Collaborate with Georgia State



In our dataset, Georgia has published more papers and engaged in more inter-institutional collaborations. Georgia has also had more authors publish in the dataset journals. Georgia had 11 different authors publish 31 articles; 3 authors were PhD students at Georgia when their articles were published. In contrast, 6 Georgia State authors published 23 articles and only 1 was a PhD student at publication time. This is not a comparison of quality between the two schools, as they are both premier institutions in the MIS field. Rather, this juxtaposition simply illustrates the effect of size: The more professors and PhD students a school has, the greater the likelihood that some long-distance collaboration will result despite the bias towards intra-institutional or short-distance collaborations[4].

So geography (certainly) and size (seemingly) of an institution affect the willingness of its researchers to engage in long distance collaborations. We then assessed whether that willingness has increased over time as the Internet and the Web have lowered the costs and barriers to distant collaborations. Consistent with other results [Rosenblat and Mobius, 2004], the table below shows that research efforts are becoming very evidently multi-institutional, rising from 1/2 of all articles in 1985-1989 to nearly 2/3 in 1996-2001.

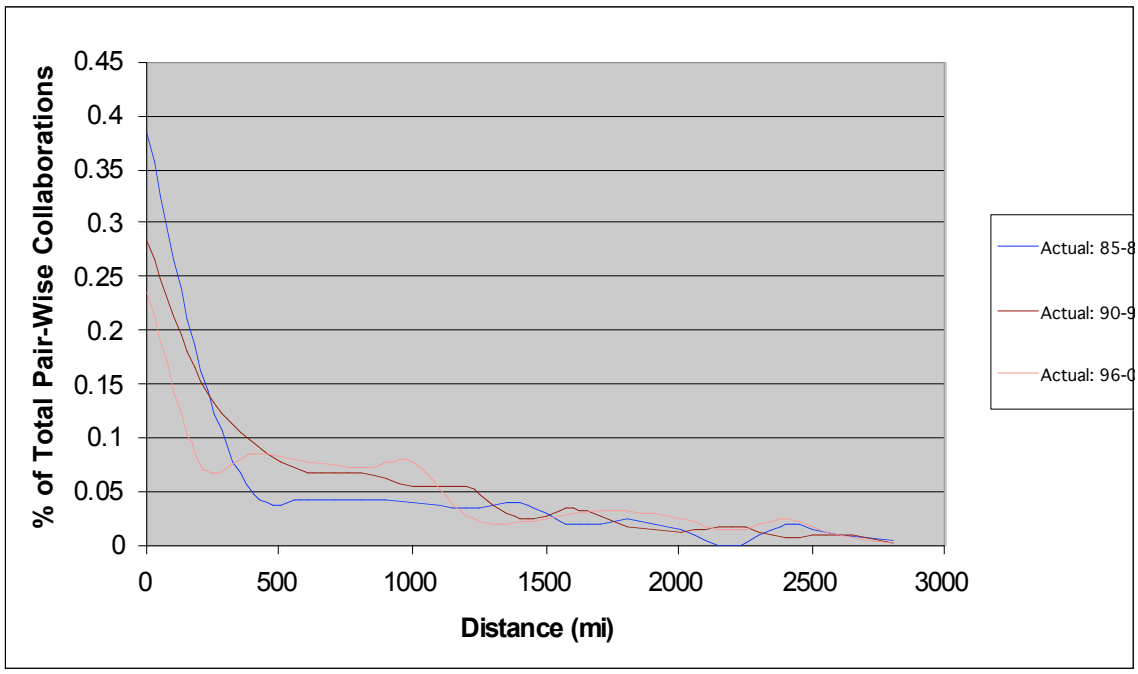
Table 1: Changes in Collaboration Patterns over Time

Period	sole author	intra-collaboration	inter-collaboration
85-89	30.4%	20.3%	49.3%
90-95	19.8%	23.1%	57.0%
96-01	21.8%	14.2%	64.0%
Grand Total	23.0%	18.8%	58.2%

[4] In fact Georgia State was ranked first in 2005 for researcher productivity. The dataset here is for a very restrictive set of journals and only for an older time period. Georgia State grew to be the largest MIS department in terms of faculty and students in the mid 1990's partly by hiring successful mid-career academics from other schools, but Georgia's MIS faculty and PhD program was more established, so the larger scale attributed to Georgia is a result of the schools' histories and the limitations of our dataset.

The next figure shows that not only has inter-institutional collaborations increased but that the distance bias also seems to have decreased. Collaboration distributions from 85-89 and 90-95 show the expected exponential decrease in collaboration as distance increases, but the distribution from 96-01 shows an initial increase in collaborations as distance increases à la the simulated, distance-unbiased distribution and then shows exponential decay. Collaboration patterns appear to be showing less distance bias.

Figure 8: Effect of Distance on Inter-Institutional Collaboration for Different Time Periods[5]



What roles has the Internet played in the mitigation of distance bias, if at all? The tails of all collaboration distributions are roughly the same; that is, collaborations beyond about 1000 miles occur roughly in the same proportions across all time periods. Therefore the effects of distance mitigation enabled by information technologies would truly be seen for collaborations of under 1000 miles. Actually the range of interest is from about 50 miles to 1000; at 0 miles there is no inter-institutional collaboration, and under 50 miles

[5] All these distributions show a heavy tail, so for ease of viewing we show a truncated view showing collaborations up to a distance of 3,000 miles.

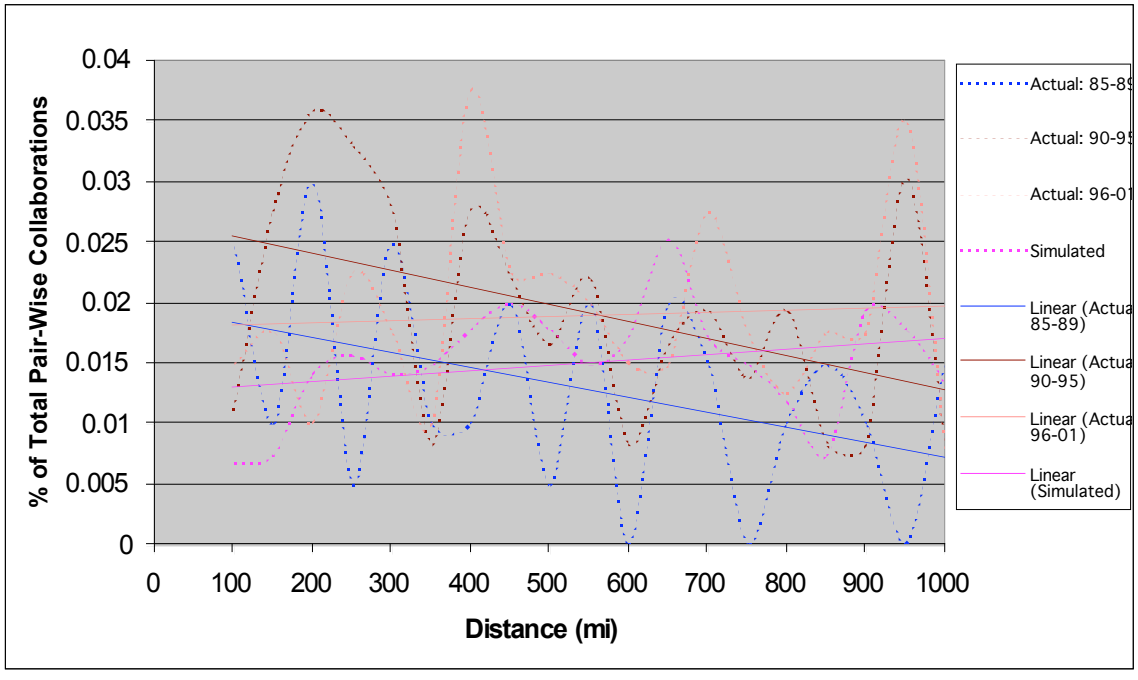
there is still a strong co-location effect. The table below shows a huge drop-off in numbers of collaborations from 0 to 1-50 miles and then a large drop from 1-50 to 51-100 miles. Thereafter changes are much less significant.

Table 2: Short-Distance Collaborations as a Percentage of all Pair-Wise Collaborations for Period

collaboration distance (in mi)	labeled as (in mi)	85-89	90-95	96-01
0	0	38.6%	28.6%	23.7%
1-50	50	9.9%	11.2%	4.2%
51-100	100	2.5%	1.1%	1.5%
101-150	150	1.0%	2.7%	1.7%
151-200	200	3.0%	3.6%	1.0%

The following chart extends this table to show patterns of collaborations labeled as from 100 to 1000 miles. Actual distributions are shown with dashed lines; a linear regression on these points to show trends is shown in solid lines. The pattern of collaboration for the simulated distribution at these distances is also shown.

Figure 9: Changes in Number of Collaborations with Distance (Shown as Linear Trend)



This result is very interesting. In the periods before 1996, the trend line shows that number of collaborations decreased as collaboration distance increased in 50 mile increments. For 96-01, the trend is pretty much flat, which is consistent with the trend for the simulated distribution of distance-unbiased collaboration. The data appears to show that researchers have started to become indifferent to collaboration distances of 200, 300, 500, or 1000 miles after 1996, and we believe that a major enabler of this indifference is likely the greater use of the Internet and the explosion of the Web.

Summary

Herein we summarize our findings:

- We found that the number of collaborations decreased exponentially with increases in collaboration distance (Figure 1) in the MIS field.

- Using a social network based community structuring algorithm, we classified institutions in this field into communities based on incidence of collaboration (Figure 2). We found that in fact the two most productive communities exhibited different collaboration patterns (Figure 3). The most productive community (Gr-GSUM) exhibited the exponential decay distribution; the second most productive (Gr-UGUT), a distribution comprised of collaborations that are not so biased by distance.

- By mapping the locations of the schools in these groups, we found that the six top schools in the first group—which was sensitive to distance in collaborations—were comprised of schools in East and West Coasts of North America where the population density and the density of schools is higher (Figure 4). The top four schools in the second group—which was less sensitive to distance—were clustered primarily in the Western or Midwestern US states where density is relatively lower (Figure 5). It appears that researchers in less densely populated areas are more willing to engage in longer distance collaborations than their counterparts in more populous areas. *Even in this age of air travel and the Internet, geography appears to still be very important in determining community structures in research even, and especially, within North America*.

- There is an anomaly with two very productive East Coast schools: Georgia and Georgia State. Whereas Georgia State belongs in the distance sensitive group, Georgia belongs in the other group. Data showed that Georgia with more published faculty and PhD students collaborated with a similar profile as Georgia State (Figure 6), but additionally collaborated with productive Western US schools (Figure 7). It is reasonable that having a larger department increases the likelihood that a few longer distance collaborations will result. *Larger institutional size may encourage more distance independent collaborations*.

- When we analyzed collaboration patterns over different time periods, we observed a noticeable difference between 1985 to 1996, and 1996 to 2001. The collaboration distribution for the later period looked more like the pattern for the simulated, distance-unbiased distribution, whereas before, the distribution showed the expected exponential decay pattern (Figure 8). Where enablers such as Internet and the Web may encourage collaborations is for collaborations between 50 to 1000 miles: Under 50 miles, co-location is a strong influence, while over 1000 miles, geography in North America means there are fewer institutions with which to collaborate. For collaborations of under 1000 miles, authors showed indifference between collaboration distances for the later period, whereas before, preference for collaboration tended to decrease as distance increased (Figure 9). *There appears to be no preference between collaborations of 200, 500, or 1000 miles now, and a major influence may be the Internet*.

Concluding Remarks

Our analysis shows that distance is still an important factor in collaborations. Incidence of collaborations decrease exponentially as the collaboration distance increases. Research communities also form based in part on close physical proximity between collaborating institutions. If however this is the effect of distance on collaboration on a large scale, on some smaller scales, its effect is being mitigated. When the data are more granularly

partitioned into various communities within the field, over different time periods, or within certain distance ranges, results under certain circumstances reflect decreased sensitivity or indifference to distance in collaboration. We believe that these effects in the smaller scale are due in large part to the Internet.

Our analysis indicates something possibly compelling. Academic co-authorship network is a scale-free network [Newman, 2001]. Our results indicate that the effect of distance on collaboration may also have had "scale free like" properties, but it appears that scale independence is being violated as a result of the Internet. For our future work, we intend to investigate this possibility. If shown to be valid, it represents a measurable piece of evidence that the Internet is changing the fundamental structure of self-organizing collaboration.

References

Newman (2004) "Fast algorithm for detecting community structure in networks," *Physical Review E* (69) no. 066133.

Katz, J. S. (1994) "Geographical Proximity and Scientific Collaboration," *Scientometrics* (31) 1, pp. 31-43.

Nagpaul, P. S. (2003) "Exploring a pseudo-regression model of transnational cooperation in science," *Scientometrics* (56) 3, pp. 403-416.

Newman, M. E. J. (2001) "Scientific collaboration networks: I. Network construction and fundamental results," *Physical Review E* (64) no. 016131.

Olson, G. M. and J. S. Olson (2000) "Distance matters," *Human-Computer Interaction* (15) 2-3, pp. 139-178.

Rosenblat, T. S. and M. M. Mobius (2004) "Getting closer or drifting apart?" *Quarterly Journal Of Economics* (119) 3, pp. 971-1009.

VanAlstyne, M. and E. Brynjolfsson (1996) "Internet - Could the Internet balkanize science?" *Science* (274) 5292, pp. 1479-1480.