Who is Our Paul Erdös? An Analysis of the Information Systems Collaboration Network

Completed Research Paper

Wallace Chipidza
Baylor University
One Bear Place #98005
Waco, TX 76798-8005
wallace_chipidza@baylor.edu

Abstract

This study examines the historical Information Systems research collaboration network. We build the network using co-authorship information in the Senior Scholar Basket of 8 journals from the publication of MISQ's first issue in April 1977 to November 2015. The different journals vary widely in their network configurations. We examine the influence of gender, temporal, and geographic homophily on co-authorship in the network. Using exponential random graph modeling on a randomly selected subset of the network, preliminary evidence suggests that ties in the IS collaboration network exhibit homophily according to gender and geography. Conversely, co-authorship seems to exhibit great diversity along the temporal dimension – researchers that graduated around the same time are not more likely to collaborate than would be expected from chance. We also reveal the current center of the IS collaboration network. Based on this center, we propose a metric to measure a researcher's connectedness within the network.

Keywords: Social network analysis, co-authorship, collaboration, homophily, exponential random graph modeling

Introduction

Researchers are not neutral observers of the research process; they exist in a temporal and sociological context. Uncovering the constitution of the Information Systems (IS) field can enhance our understanding of how the field creates scientific knowledge. In other words, IS research is a community of practice in which knowledge creation is a product of social interaction (Gallivan and Ahuja 2015). Indeed, calls for recording and preservation of the history of IS research have increased in recent years (Zhang 2015). Situating the field in its proper sociological context allows the revelation of latent biases which might impact the trajectory of parts of our research.

This study reveals the historical IS collaboration network (C_{IS}) and articulates its properties. Every academic discipline has a collaboration network that can be created by linking all co-authors. This network can be viewed as an artifact of the discipline, and can be useful in representing the history of the discipline. Such a construction is consistent with Mason et al. (1997)'s articulation that "if MIS is to enjoy the theoretical and professional recognition that academic maturity bestows on a discipline… MIS professionals must begin also to record and examine its history" (pp. 258). Although Mason et al. (1997)'s study focused mainly on the need to understand the history of organizations in the context of technological use, their argument equally applies to the IS research field as a whole.

As research in IS escapes nascence, enough structural information about its network is being accumulated. Collaboration in a research field is positively associated with scholarly output, which forms

the basis for reward assessment in academia (Gallivan and Ahuja 2015). The connections wrought by collaboration weave into a complex network where researchers influence each other in the knowledge production process (Xu et al. 2014). However, to date, there has been little attention towards understanding the dynamics of collaboration in IS research (Gallivan and Ahuja 2015, pp 981). To demonstrate the utility of the collaboration network, we invoke the example of Paul Erdös - the highly productive Hungarian mathematician who is widely cited as the center of the math research universe (Grossman 2002). Mathematicians trace their connections to Paul Erdös through use of the Erdös number, a measure of the collaborative distance between any given author and Paul Erdös (De Castro and Grossman 1999). In order to characterize the field of IS, capturing characteristics of the network such as its centers of influence is necessary so that our luminaries are similarly recognized. The network allows us to view the most connected researchers in our field (Cuellar et al. 2016).

We aim to analyze the historical IS research network to uncover its interesting characteristics and in the process, reveal the IS equivalent of Paul Erdös. Further, we shall compare the different sub networks corresponding to each journal in the sample. To construct the network, we choose a starting point of 1977 – the year of *MIS Quarterly*'s first issue, and track collaboration information in the senior scholars' Basket of 8 journals up until 2015. We create the network with information from the following journals: *Management Information Systems Quarterly* (MISQ), *Information Systems Research* (ISR), *Journal of Management Information Systems* (JMIS), *Journal of the Association for Information Systems* (JAIS), *European Journal of Information Systems* (EJIS), *Information Systems Journal* (ISJ), *Journal of Strategic Information Systems* (JSIS), and *Journal of Information Technology* (JIT) – a total of about 6,000 articles.

Several questions shall be answered in this study: first, what are the characteristics of the IS collaboration network? Second, how do the sub-networks corresponding to the different journals differ, if at all? Third, how do author characteristics determine tie formation? The first two analyses will be conducted on the full network of 5670 nodes and 10303 edges, and the third will be conducted on a randomly selected sub-network of 208 nodes and 110 edges.

The rest of the paper is sequentially organized as follows: we present a review of related work and outline the theoretical underpinning of the study. Afterwards, we detail the methods of data collection and analysis before presenting the results, discussing our findings, and concluding with a summary.

Literature Review

Several studies have focused on structural aspects of the IS collaboration network (Xu et al. 2014; Zhai et al. 2014). Perhaps the most comprehensive description of the IS network was conducted by Xu et al. (2014). To investigate their hypotheses, Xu et al. examined co-authorship data from the Basket of 6 journals, covering the period between 1980 and 2012. The study found that, over time, IS has acquired social capital and become more connected. The focus of the study was on describing the structural evolution of the network. Network structure results from tie formation. Therefore, to build on Xu et al (2014)'s findings, our study tries to understand the antecedents to tie formation – we explore the effects of author characteristics on tie formation in the network. In addition, it is highly likely that expanding the journal set to incorporate at least the Basket of 8 journals and the period after 2012 will generate new information about the IS collaboration network.

The IS field has traditionally been plagued by underrepresentation of women and minorities (Coder et al. 2009). This underrepresentation may extend to the IS academe as well. Not surprisingly therefore, past IS research has investigated homophily in collaboration according to gender. Gallivan and Ahuja (2015) found that collaboration in five mainstream IS journals exhibits homophily according to gender and institution of graduation. Two opportunities for further research are presented. First, Gallivan and Ahuja (2015) did not explore whether gender homophily is stronger in male or female researchers. Such a determination would be useful in illuminating whether women find it difficult to find co-authors, relative to men. Second, the set of journals covered in Gallivan and Ahuja (2015)'s study excluded 3 journals in the Senior Scholar Basket of 8 journals and only covered the period between 1999 and 2005. The authors acknowledged these limitations of their study, and encouraged future research to examine whether their findings generalize across more journals and across a longer time period. In that spirit, we expand the number of journals to 8 and the period of examination from 6 years to 38 years. Third, in addition to

gender and geographic homophily, we also examine the influence of field tenure (the length of time since a researcher graduated with a PhD) on collaboration in IS research.

Another common theme in research network studies concerns metrics that capture a researcher's potential and/or past productivity. To that end, researchers have proposed various measures such as the h-index and the number of publications in selected journals (Lowry et al. 2013). Such studies typically find problems within the status quo; hence they propose new metrics of assessing scholarly influence. Most recently, Cuellar et al. (2016) proposed that scholarly capital should be measured as an aggregation of three measures: ideation, venue representation, and connectedness. The latter would be measured by how close a researcher is to influential researchers within the field. Connectedness is an important measure of scholarly influence because it predicts a researcher's scholarly output (Lowry et al. 2013). Our study can further illuminate how a researcher's connectedness within the IS network can be easily measured.

Theoretical Development

There are at least two competing influences on tie formation in a network. The first is heterophily, meaning that people with different characteristics would form collaboration links with a higher probability than would be expected from chance (Currarini et al. 2009). For example, some researchers might elect to collaborate with people that possess core competencies that they lack. If that tendency to collaborate on the basis of resource complementarity is sufficiently widespread in the network, then heterophily might explain tie formation in the network. The desire to complement resources might then dwarf other factors that potentially affect collaboration in the network, such as gender, race, or geographic location.

The second competing influence is homophily, which is the tendency for people to form connections with others of similar backgrounds (Currarini et al. 2009). Consistent with the idiom "birds of a feather flock together," ties in any real network are likely to form among people of similar characteristics (Lazarsfeld et al. 1954; McPherson et al. 2001). This has been termed the similarity-attraction hypothesis – interaction is more likely to occur among people with similar traits (Yuan and Gay 2006). Another explanation for homophily is the self-categorization principle – the tendency for individuals to place themselves and others in categories according to characteristics such as gender, age, and race (Turner et al. 1987). These categories allow individuals to categorize others as similar or dissimilar, which forms the basis for homophilous ties. Homophily may be explained by information flow – people with similar characteristics are likely to communicate more easily than people with different characteristics (Egorov et al. 2010). Therefore, homophily predicts that there are more homogeneous (male-to-male or female-to-female) co-authorship ties than heterogeneous (female-to-male or male-to-female) ties in the IS field (H1A).

The extent to which women form homophilous ties is likely to differ from the extent to which men form homophilous ties in the IS field. This is a consequence of self-categorization – on average, an individual belonging to a majority group places him/herself in a larger category compared to an individual belonging to a minority group. As shown by Currarini et al. (2009), larger groups tend to exhibit higher homophily in tie formation than small groups. The proportion of female IS researchers is likely low (Gallivan and Ahuja 2015). To illustrate the hypothesized difference in homophily according to gender, consider an extreme case in which only one woman exists in the IS field. Because the field has an extreme male majority, the woman has no choice but to collaborate with a man (assuming that collaboration is preferred to sole authorship). It is only if the number of women in the field is continually increased that, at a certain point, women will find it feasible to preferentially collaborate with other women. Because estimates for the composition of female researchers range from 20 to 30% (Gallivan and Ahuja 2015), we expect that women will demonstrate less homophily (even heterophily) than men in their choice of co-authors in the IS field (H1B). On the other hand, men have an extreme majority in the IS field; hence, it is likely that men will demonstrate homophily in their choice of co-author.

In addition, large groups tend to form more ties per capita than small groups (Currarini et al. 2009). If homophily is the primary mechanism through which collaboration occurs, then according to the self-categorization principle, men have more potential collaborators in their category on average, compared to women. Therefore, we expect that male IS researchers will, on average, collaborate with more people than female researchers (H1C). These considerations lead us to propose that, in the historical IS network:

H1A: Ties are more likely to form between people of the same gender than between people of different genders.

H1B: Ties are more likely to form between male researchers than between female researchers.

H1C: Male researchers will have more ties on average than female researchers.

Further, there is likely to be homophily according to time period of PhD degree attainment (temporal homophily): network ties are more likely to form among researchers that graduated in the same time period than among people that graduated in different time periods (H2A). For this study, we divided the time periods based on decades: pre-1980s, 1980s, 1990s, 2000s, and 2010s. The self-categorization effect predicts that researchers that graduated around the same time would place others and themselves in one category because of common experiences. For example, they might have met in the same PhD program, at workshops, conference presentations, and/or doctoral consortia. They are likely to have had more opportunities for contact than researchers that graduated in different time periods. These are examples of social structuring of activities, where similar people are brought into contact more frequently than one would expect from chance (Feld 1982). Further, the more time a researcher is part of the field, the greater the human capital he or she acquires. As a result, a researcher with an older PhD will have more opportunities to collaborate with other people than a researcher with a young PhD (H2B). These considerations lead us to propose that, in the historical IS network:

H2A: Ties are more likely to form between people that graduated in the same period of time than between people that graduated in different periods.

H2B: Researchers with older PhDs are likely to have more ties than researchers with young PhDs.

There is likely to be homophily according to location of PhD-granting institution: the formation of a network tie is likely to be affected by geographical distance (geographic homophily). PhD graduates are more likely to stay in the country from which they graduate for a variety of reasons. First, institutions in the same country are likely to have similar recruitment standards and expectations for tenure. Second, institutions are likely to have existing recruitment relationships with institutions located in the same geographical region. For example, students that graduate from New York University usually end up at the Massachusetts Institute of Technology, and the opposite is true (Gallivan and Ahuja 2015). Third, the costs of relocating to a different geographical region are higher than the cost of staying within the same region. Because of the preceding reasons, researchers will more likely collaborate with people with whom they have established relationships, and these people are likely to have been in the same PhD program (as students or faculty). This enables us to capture collaboration between faculty and their PhD students, as opposed to just collaboration between students of the same PhD program (see Gallivan and Ahuja 2015). For this study, a geographical region was defined as a continent (Africa, Asia, Australasia, Europe, North America, and South America). These arguments lead us to propose that:

H3: Ties are more likely to form among researchers that graduated in the same geographical region than among researchers from different geographical regions.

Table 1 below summarizes our study hypotheses.

	Hypothesis
H1A	Ties are more likely to form between people of the same gender than between people of different genders.
H1B	Ties are more likely to form between male researchers than between female researchers.
H1C	Male researchers will have more ties on average than female researchers.
H2A	Ties are more likely to form between people that graduated in the same period of time than between people that graduated in different periods.
H2B	Researchers with older PhDs are likely to have more ties than researchers with young PhDs.
H3	Ties are more likely to form among researchers that graduated in the same geographical region than among researchers from different geographical regions.

Table 1: Research Hypotheses

Method

In order to create the collaboration network, we needed details of a representative sample of co-authorship ties in the field of Information Systems. The Senior Scholars' Basket of 8[1] journals is a widely accepted sample of IS journals that is frequently employed as a representative journal collection in scientometrics and in literature reviews (Li and Karahanna 2015; Zhai et al. 2014). As such, we selected the Basket of 8 journals as the representative sample of the IS field. Technical details of our data collection and parsing process are included in the Appendix (page 15).

Exponential Random Graph Modeling

Until recently, most work in social network analysis focused on describing certain quantifiable measures of the network. For example, transitivity has generally been used as a proxy for network connectedness, with higher levels of transitivity signaling better connectedness (Wyatt et al. 2008). In the same vein, the extent to which nodes mix according to some criterion has been measured using the assortativity measure (Handcock et al. 2008). While these measures are useful in describing the network, they do not explain or predict how the network forms. Exponential random graph modeling (ERGM) is useful because it enables explaining or predicting the probability of a network tie based on some attributes of the nodes (Handcock et al. 2008). Therefore, ERGM enables the researcher to model a network and hence enact Gregor (2006)'s Type IV theoretical contributions using network analysis. Specifically, we can employ ERGM to predict the probability of a tie between any two nodes according to node characteristics.

ERGM is analogous to a binary logistic regression model in that the outcome variable is binary: whether a network tie is formed or not between any pair of nodes (Wasserman and Pattison 1996). The social network is considered as a set of random variables, and each potential edge in the network is associated with one variable (Wyatt et al. 2008). ERGM allows us to assign probabilities to a tie between any two nodes based on some property. The following equation describes the model:

$$p(Y = y|\theta) = \frac{1}{Z} e^{\theta^T \phi(y)}$$

where Y is the set of variables representing edges in the network, y is the observed network adjacency matrix, Z is a normalizing constant that ensures that the probabilities stay within the 0 to 1 range and the

[1] "Senior Scholars' Basket of Journals," Association for Information Systems

https://aisnet.org/?SeniorScholarBasket

probabilities across all networks sum up to 1, θ represents a vector of weights to be learned, and ϕ represents the feature functions defined on y (Harris 2013; Wyatt et al. 2008).

Large networks are usually not feasible to examine or visualize in their entirety. As a result, a sampling strategy that selects a representative sub network of the network can be useful in demonstrating aspects of the network. The collaboration network for IS research, C_{IS}, has 5670 nodes and 10303 edges, and it is not particularly large compared to other real life networks such as older academic disciplines, Facebook, and the Internet (Barabási and Bonabeau 2003). However, we can only automate the collection of collaboration information from journals' tables of contents; demographic information such as gender and year of graduation for each author has to be manually collected. Because the effort required to manually collect the information is huge, we found it prudent to select a subsample of the network on which to perform our analyses. A variety of sampling strategies exist; these strategies include random node sampling, random edge sampling (RES), random walk sampling, and random node neighbor sampling (Leskovec and Faloutsos 2006).

For this study, we employed random edge sampling because the structure of the resulting sub network is well understood. The main drawbacks of random edge sampling are that the resulting sub network will be biased towards high-degree nodes, and the resulting sub network will be sparsely connected and thus not respect the underlying community structure (RES is still better at respecting network structure than random node sampling) (Leskovec and Faloutsos 2006). The reason for the former is that, by definition, high degree nodes have many more edges per node than the average node. As a result, the probability that a high-degree node is part of a randomly selected edge list is quite high. C_{IS} has a power law distribution, and new members join the network through preferential attachment (Xu et al. 2014). Therefore, understanding the homophily in an RES sub network can be useful in revealing the behavior of the most popular nodes in the network. The main advantage of RES sampling is that it is relatively simple. To conduct ERGM on C_{IS}, we randomly selected 110 ties from the consolidated edge list of C_{IS}. We carried out Google searches to discover demographic information for the randomly selected 208 researchers. The information collected were gender, year of PhD award (where applicable), and geographical location of PhD awarding institution. The above three attributes are fixed, and hence are the most appropriate attributes on which to anchor the ERGM.

Results

In presenting the results of our analysis, we distinguish between descriptive and modeling statistics. Descriptive statistics, particularly the betweenness measure, are useful in revealing the current center of C_{IS}. Modeling statistics reveal how homophily influences edge formation in the network.

Descriptive Statistics

Up until November 2015, a total of about 5,500 research articles had been published in the Basket of 8 journals (we removed book reviews and editorials from the data set). JMIS has the highest number of publications, and ISJ the lowest. About 22% of these papers were written by one author, and 78% had at least two authors. This suggests that collaboration is preferred to sole authorship in IS research. Overall, JAIS has the highest number of authors per research article, and EJIS the lowest (Table 2). Compared to the overall mean number of authors per article (2.32), 4 journals have a higher statistic. It is interesting to note that three of these journals also form the top tier of journals in the rankings provided by AIS[2].

[2] https://aisnet.org/general/custom.asp?page=JournalRankings

	Number of Publications	Single Author Papers	Multiple Author Papers	Total Number of Authors	Number of Authors per Article
MISQ	1087	211	876	2538	2.33
ISR	671	89	582	1703	2.54
JMIS	1154	183	971	2889	2.50
JAIS	410	63	347	1077	2.63
ISJ	378	107	271	815	2.16
EJIS	812	281	531	1671	2.06
JIT	552	152	400	1156	2.09
JSIS	499	165	334	1045	2.09
Combined IS	5563	1251	4312	12894	2.32

Table 2: Collaboration Statistics for Basket of 8 Journals

On average, each author in the IS research network has collaborated with 4.19 authors, with a standard deviation of 6.79. Figure 2 shows the degree distribution of C_{IS}, compared to the degree distribution of a random network of the same size and edge density. The distribution of node degrees follows a power law distribution; hence, C_{IS} resembles other real world networks. This means that there are few highly popular nodes and a majority lowly popular nodes. A randomly generated network, on the other hand, will have a normal distribution of ties per node. The power law degree distribution implies that C_{IS} is scale free – the network grows with time; as such, there are opportunities for new members to join the network. Further, the network grows with preferential attachment – new members join the network by collaborating with more established researchers. The end result is that already popular nodes are more likely to acquire more links; hence obeying the "rich get richer" model (Barabási and Bonabeau 2003).

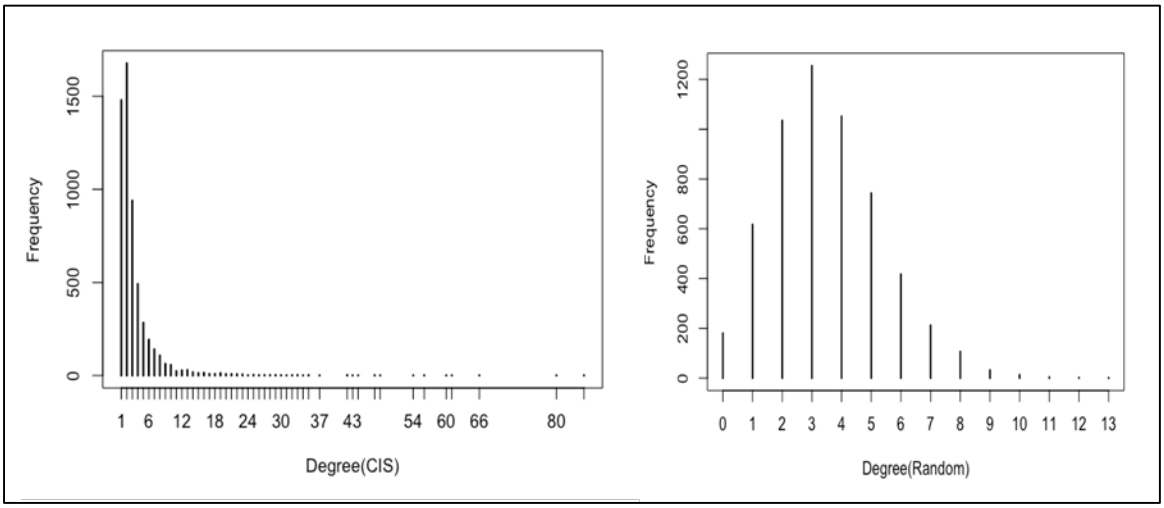

Figure 2: Degree Distributions of C_{IS} vs Randomly Generated Graph of Comparable Size and Edge Density

Table 3 shows the different network metrics for the journals in the Basket of 8. MISQ, EJIS, ISR, and JMIS have high numbers of authors. MISQ has the largest connected component, with 781 vertices. It also has the longest diameter – 17 – which is the maximum shortest path from one author to another. On

average, in the MISQ network's largest component, an author has to take about 7 steps to find another author. The diameter for C_{IS} is 20, and the average shortest path length is 7.38.

Journal	Number of Nodes	Number of Edges	Proportion of Largest Component	Diameter	Average Path Length
MISQ	1479	2178	0.52	17	6.66
ISR	967	1515	0.56	15	5.72
JMIS	1607	2475	0.47	13	4.48
JAIS	786	1091	0.24	9	3.19
EJIS	1052	1244	0.15	6	1.85
ISJ	613	779	0.06	4	1.55
JSIS	739	805	0.05	7	2.59
JIT	933	991	0.14	7	2.85
COMBINED IS	5670	10303	0.67	20	7.38

Table 3: Network Metrics for Journals in the Basket of 8

Figure 3 shows visualizations of the collaboration networks belonging to the different sub-journals. MISQ, ISR and JMIS have relatively large, discernible cores. Because the above journals consistently occupy high ranks on business journal rankings (Lowry et al. 2013), a stable core seems to be a requirement for (or perhaps a consequence of) journal success. On the other hand, ISJ and JSIS are sparsely populated and their cores are very small, if they exist at all. In addition, C_{IS} has an even greater core, comprising 67% of the network. This represents an increase from the 65% reported by (Xu et al. 2014) in describing the 1980 – 2012 IS network. This strongly supports the argument that C_{IS} is becoming more connected with time.

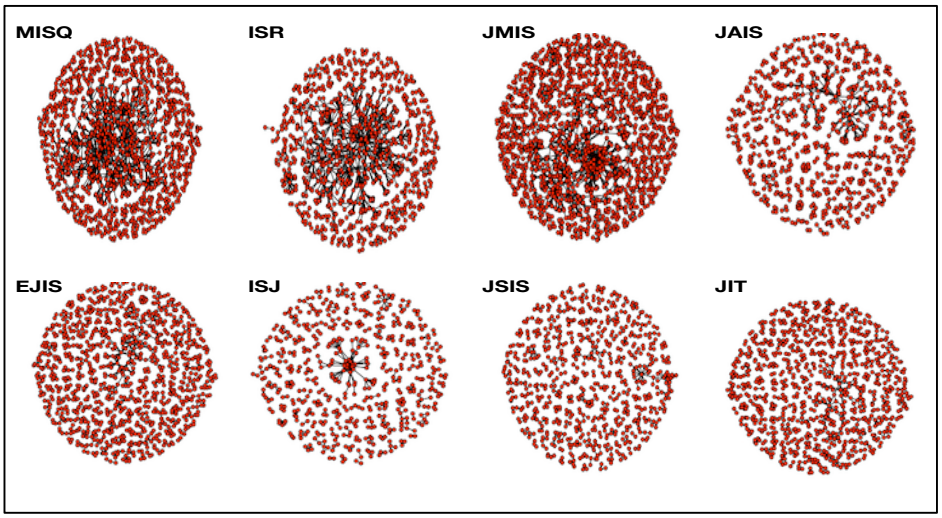

Figure 3: Network Configurations for the Basket of 8 Journals

C_{IS}, the combined network of all the journal networks, is more connected than any of the individual networks. It has a dense, stable core (Figure 4), meaning that the network is robust and can survive the loss of its key nodes. Such existential debates as have dominated the discourse in the past, are perhaps less relevant now. There is a growing recognition that IS is a legitimate field requiring little justification of its existence. The robustness of C_{IS} is an empirical manifestation of its legitimacy.

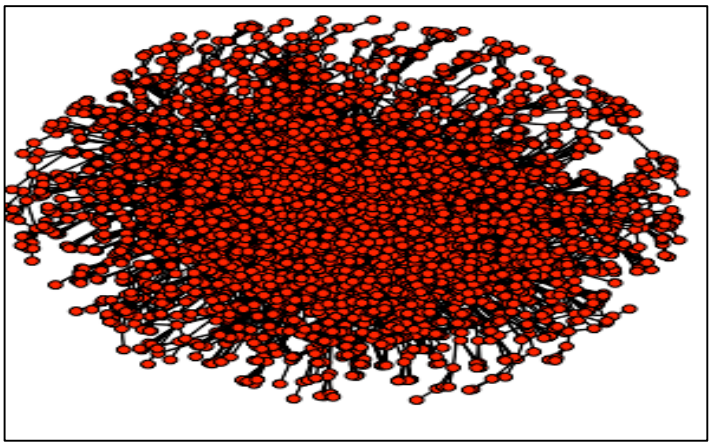

Figure 4: Network Configuration for C_{IS}

Table 4 shows the ten most central IS researchers according to betweenness centrality – a measure of the extent to which a network node lies between other nodes (Xu et al. 2014). Researchers with high betweenness scores act as brokers of knowledge transfer within the network (Xu et al. 2014). The majority of the most central IS researchers have served as editors-in-chief of our elite journals. For example, Izak Benbasat and Ritu Agarwal served as editors-in-chief for ISR, Kalle Lyytinen served as editor-in-chief for JAIS, and Arun Rai is the current editor-in-chief for MISQ, a post that Detmar Straub has occupied in the past. This suggests that betweenness centrality is valuable and informative. It is important to emphasize that betweenness centrality is not a measure of scholarly impact or productivity, rather, it is a measure of how connected a researcher is in the network. Some highly productive researchers may prefer sole authorship to collaboration, or they may publish in other prestigious journals that are not in the Basket of 8 e.g. *Management Science* or *Academy of Management Journal*; hence productive scholars will not necessarily have high centrality scores in C_{IS}. Nevertheless, as we show below, highly productive researchers are likely to have higher than average betweenness scores.

Name	Affiliated Institution	Betweenness Centrality
Kalle Lyytinen	Case Western Reserve (USA)	597033.45
Izak Benbasat	University of British Columbia (Canada)	519542.69
Alan Dennis	Indiana University (USA)	453721.58
Paul Pavlou	Temple University (USA)	407611.29
Ritu Agarwal	University of Maryland (USA)	388043.36
Detmar Straub	Temple University (USA)	338316.95
Arun Rai	Georgia State University (USA)	326785.77
Gordon Davis	University of Minnesota (USA)	309470.54
Jay Nunamaker	University of Arizona (USA)	298837.53

Table 4: Highly Connected IS Researchers According to Betweenness Centrality

Kalle Lyytinen of Case Western Reserve University has the highest betweenness centrality score in C_{IS}, meaning that his removal from the network would penalize everyone else's average shortest path length the highest. By the betweenness measure, Lyytinen qualifies as the current center of the historical C_{IS}, and he is therefore the IS equivalent of Paul Erdös. Lyytinen publishes prolifically both in North American and European journals. As described earlier, the Erdös number of an author in Math research is the collaboration distance between the author and Paul Erdös. Similarly, we define the Lyytinen number (LN) of an author in IS research as the collaboration distance between an author and Kalle Lyytinen.

An author that has collaborated with Lyytinen has an LN of 1; an author that has not directly collaborated with Lyytinen but has collaborated with a co-author of Lyytinen has an LN of 2, and so on. The minimum LN is 0, and corresponds to Lyytinen's collaboration distance from himself. The maximum Lyytinen number in the largest component of C_{IS} is 10, which is also Lyytinen's eccentricity. As we stated earlier, the largest component of C_{IS} contains 67% of the researchers, hence, 67% of all researchers in C_{IS} have an LN. The median LN in the largest component of C_{IS} is 4 (mean = 3.84, sd = 1.33). For the top researchers in the IS field (see footnote below), the median LN is 3 (mean = 2.85, sd = 1.14). Figure 5 conveys this information. The differences between the two group means is statistically significant (F = 68.53, df = 1, p = 0.000). These statistics suggest that top IS researchers have low LNs. The same phenomenon is evident in Math research, where successful mathematicians tend to have lower Erdös numbers than the average (De Castro and Grossman 1999).

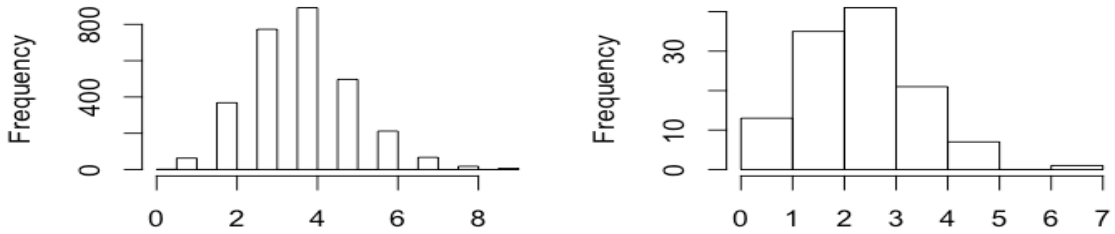

Figure 5: Distribution of Lyytinen Numbers for Authors in C_{IS} (left histogram) and Top IS Researchers (right histogram)[3]

Alternative measures of network influence also exist, such as degree, closeness, and eigenvector centrality (Bonacich 2007). The latter measure is particularly useful because it captures the importance of a node's connections – researchers with high eigenvector centrality are connected with other important researchers (Bonacich 2007). Table 5 shows different journal combinations and the researcher with the highest betweenness and eigenvector centrality score across each combination. From the table, one can convincingly argue that Izak Benbasat and Jay Nunamaker consistently occupy the most influential positions in the historical IS research network. However, in identifying the center, we were guided by the strict parallel of the math research network, which employs the betweenness measure to locate Erdös as the center (De Castro and Grossman 1999; Grossman 2002).

Journal Combination	Highest Betweenness Centrality Score	Highest Eigenvector Centrality Score
Basket of 8	Kalle Lyytinen	Jay Nunamaker
MISQ + ISR	Izak Benbasat	Detmar Straub
MISQ + ISR + JMIS	Izak Benbasat	Jay Nunamaker
MISQ + ISR + JMIS + JAIS	Izak Benbasat	Jay Nunamaker
Basket of 6	Izak Benbasat	Jay Nunamaker

Table 5: Highly Connected IS Researchers Across Journal Combinations

Modeling Statistics

The null model represents the baseline model which we can compare against after accounting for the influence of node attributes on the formation of a tie in the model (Harris 2013). Akaike's "An Information Criterion" (AIC) is a formula that is used to compare models that are fitted according to maximum likelihood to the same data (Akaike 1974). The smaller the AIC, the better the model. The italicized text is *R* code. The edges term demonstrates the propensity for edges to form in the network,

[3] We employed the H-Index list maintained at the University of Arizona

https://ai.arizona.edu/sites/ai/files/MIS510/h-index-2015-04.pdf

this is typically low in real world networks (Harris 2013); we can conclude from the negative coefficient that the structure of the network is characterized by a low probability of edge formation. The following *R* code builds the baseline model:

> *ergm(nrelations ~ edges)*

	Coefficient	**Std Error**	**p-value**
edges	-5.27151	0.09558	.000***

Table 6: Null Model for RES C_{IS} Sub-Network (AIC = 1382)

Having created the baseline model, we examined whether node attributes have any influence on the likelihood of tie formation. We tested the homophily effects in the model, in order to test hypotheses H1A, H2A, and H3. This required the addition of gender, geographical region of PhD program, and period of PhD graduation researcher attributes to the model as main effects. Below are the null and alternative hypotheses for the main effects of researcher gender on the likelihood of co-authorship:

H_0 (gender homophily): For any two researchers, there is no relationship between their genders and the likelihood that they will collaborate.

H_A (gender homophily): For any two researchers, there is a relationship between their genders and the likelihood that they will collaborate.

H_0 (temporal homophily): For any two researchers, there is no relationship between the times that they attained their PhDs and the likelihood that they will collaborate.

H_A (temporal homophily): For any two researchers, there is a relationship between the times that they attained their PhDs and the likelihood that they will collaborate.

H_0 (geographic homophily): For any two researchers, there is no relationship between the geographical regions in which they attained their PhDs and the likelihood that they will collaborate.

H_A (geographic homophily): For any two researchers, there is a relationship between the geographic regions that they attained their PhDs and the likelihood that they will collaborate.

We use the edges term, gender, geographic, and temporal homophily terms as predictors of the probability that a tie will form. The homophily terms capture the degree to which nodes of similar characteristics tend to form edges, over or below what would be expected from chance. The *nodematch* command tests for homophily. The results in Table 7 show positive and significant homophily coefficients for gender and geographical region of PhD program, but not for the period of PhD attainment in the subnetwork of 208 authors and 110 edges. Therefore, we reject the null hypotheses for the gender and geographic attributes, but we fail to reject the null hypothesis for the temporal attribute. In addition, the AIC decreased after adding the homophily terms (compared to the AIC for the null model), indicating that the homophily model fits the observed data better than the null model. Hence, we find empirical support for H1A and H3, but not for H2A.

> *ergm(nrelations ~ edges + nodematch('Gender') + nodematch('Region') + nodematch('GradPeriod'))*

	Coefficient	**Std Error**	**p-value**
edges	-6.3709	0.2728	.000***
nodematch.Gender	0.4577	0.2281	.0448*
nodematch.Region	1.0878	0.2279	.000***
nodematch.GradPeriod	0.187	0.2149	.3842

Table 7: Main Effects Model for RES C_{IS} Sub-Network (AIC = 1356)

It's worth noting that we conducted journal-by-journal analysis of homophily influences within the network, and found no significant homophily influences in any of the journals. As we collect more data, such influences are likely to become more visible, if they exist.

We ran a multiple regression test to determine the effect of an author's gender and his or her tenure of PhD on the degree centrality of an author. The degree centrality of authors varied according to the tenure of their PhDs (β = 0.41, p = .000), but did not vary according to the gender of an author. Further, the tenure of the PhD explained 10% of the variance in degree centrality. Specifically, the longer the tenure of the PhD, the greater the number of collaborators. These findings provide support for H2B, but not for H1C.

To test H1B, we ran the *mixingmatrix* command in *R* so that we could visualize the relative proportions of MM versus FF versus FM or MF edges in the RES sub network. Table 8 shows that in the sampled sub-network, MM edges comprised the majority of co-authorship links (70%), followed by heterophilous links (23%), and then FF links (7%).

> *mixingmatrix(nrelations, 'Gender')*

Right side author / Left side author	Male	Female
Male	77	25
Female	25	8

Table 8: Collaboration Combinations According to Gender

The results show that male researchers are more likely to collaborate with other male researchers than with female researchers (χ^2 = 26.51, df = 1, p = .000). On the other hand, female researchers are more likely to collaborate with male researchers than with female researchers (χ^2 = 8.76, df = 1 = .003). These findings lend empirical support towards H1B. Table 9 summarizes our findings.

	Hypothesis	**Result**
H1A	Ties are more likely to form between people of the same gender than between people of different genders.	Supported
H1B	Ties are more likely to form between male researchers than between female researchers.	Supported
H1C	Male researchers will have more ties on average than female researchers.	Not Supported
H2A	Ties are likely to form between people that graduated in the same period of time than between people that graduated in different periods	Not Supported
H2B	Researchers with older PhDs are likely to have more ties than researchers with young PhDs.	Supported
H3	Ties are more likely to form among researchers located in the same geographical region than among researchers from different geographical regions	Supported

Table 9: Research Outcomes

Finally, we calculated odds ratios for the homophily effects in the RES network. The odds ratios in Table 10 show that two researchers of the same gender are 1.58 times more likely to collaborate than two researchers of different genders (95% CI: 1.01 to 2.47). Further, two researchers that graduated from the same geographical region are 3 times more likely to collaborate than two researchers that graduated from different geographical regions (95% CI: 1.9 to 4.64). On the other hand, two researchers that graduated at around the same time are *not* any more likely to collaborate than two researchers that graduated in different eras.

	Lower	Odds Ratio	Upper
nodematch.Gender	1.01	1.58	2.47
nodematch.Region	1.9	2.97	4.64
nodematch.GradPeriod	0.79	1.21	1.84

Table 10: Odds Ratios for Homophily Model

Discussion

This study makes several contributions towards our understanding of collaboration in the IS research field. First, we assess the role of homophily in determining collaboration in the field. Past research has examined gender and geographic homophily (Gallivan and Ahuja 2015). According to the self-categorization principle, researchers categorize themselves across different measures. Hence, we assessed geographic homophily differently than in previous studies. Whereas Gallivan and Ahuja (2015) defined geographic homophily as attending the same PhD program, we broadened the definition to be the region of PhD program. The difference between the two measures is subtle, but it allows us to capture collaborations between faculty and their (former) students (see explanation in the Theory section); this collaboration would not be captured by Gallivan and Ahuja's definition of geographic homophily. This allowed us to examine the prevalence of cross-continental collaboration in IS, versus intra-continental collaboration.

Further, we examined the role of temporal homophily in determining collaboration and found that IS researchers generally do not self-categorize along this dimension; we found no support for temporal homophily. This made us wonder whether the opposite effect (heterophily) was prevalent across this dimension, as opposed to homophily. In other words, do senior faculty collaborate with junior faculty or their PhD students more than would be expected from chance, for example? Instead of the *nodematch* command, we used the *absdiff* command, which measured whether, for any two researchers, the likelihood of collaboration increased with a greater difference between the ages of their PhDs. This relationship was not significant. This suggests that collaboration among senior faculty and junior faculty is just as prevalent as collaboration among senior and senior faculty, for example. Collaboration teams therefore exhibit great diversity along this measure – a three-member team might have a PhD student, a junior faculty researcher and a senior faculty researcher. This finding shows that there is diversity in collaboration according to field tenure, which bodes well for creativity and innovation in IS research (Joshi and Roh 2009). Conversely, we found support for gender and geographic homophily. With the rising prominence of cloud-enabled research collaboration tools such as Dropbox and Zotero, it is possible that geographic homophily will disappear with time, because of depleting necessity for physical co-presence when co-authoring research articles. Individual researchers might reflect on whether they should consider collaboration with researchers based in other locations, especially when they meet at conferences and workshops. More heterophilous research according to gender and geography would also result in higher levels of idea sharing, leading to higher quality research and increased focus on understudied areas of the field. For example, the topic of ICT4D has received limited focus in the Basket of 8 journals, possibly because sub-Saharan Africa is barely represented in the IS field.

Our study differed from previous collaboration research in two different ways. We examined collaboration across all journals in the Basket of 8 journals, and our analysis spanned across the period from 1977 to 2015. This enables a more holistic examination of the current historical network.

Although multiple studies recognize the importance of a researcher's connectedness in the research network (Cuellar et al. 2016; Lowry et al. 2013), none of them actually provide an easily accessible measure. Our revelation of Kalle Lyytinen as the current center of C_{IS} provides an easy metric of determining a researcher's embeddedness within the network. The LN for a researcher is a proxy for his/her distance from the center of the research network. For any given researcher, his/her LN is easily discoverable through such tools as Microsoft Academic Research[4] and DBLP[5]. We showed that the most

[4] https://academic.microsoft.com/

productive researchers in IS have on average, lower LNs than the general population of researchers. In other words, top researchers are likely to be close to the center of C_{IS}. The productivity and standing of a researcher's advisor and co-authors are likely to influence his/her impact on IS research (Lowry et al. 2013). The Lyytinen Number may thus be used to assess how influential the collaborators and advisors are – a lower average LN for the author's advisor and co-authors suggests that a researcher possesses high potential to impact the field. In this regard, the LN may prove useful for PhD students seeking academic advisors and for hiring committees as a piece of additional information when making their assessments.

Last, the Basket of 8 journals show a wide variety in network configuration. Compared to other journals, MISQ (52%) and ISR (56%) have relatively large maximally connected components; hence they each have stable cores. These two journals are widely viewed as the top IS journals on such lists as the Financial Times and UT-Dallas (Lowry et al. 2013). Stable cores accumulate social capital for the journals. The larger the network, the more value it has (Metcalfe 1995), therefore journals with large maximally connected components are likely to be more valuable. Second, higher than average collaboration levels are found in North American-based journals. This suggests that collaboration cultures vary across geographical regions – with North American-based journals having higher levels of collaboration and larger maximally connected components. Journals with lower levels of collaboration might want to foster higher levels of collaboration in order to increase their social capitals. An example of such a venture is ISJ's requirement for each research team submitting to its special issue on ICT4D to have at least one author from a developing country[6]. As more IS journals accumulate social capital, so does the IS research field as a whole.

Limitations and Future Research

The results of our study must be considered in the context of its several limitations. First, the network contained 10,000 edges; we ran our analysis on 110 of these edges because manual collection of researchers' demographic data takes considerable effort and time. We can increase our confidence in the results with a much larger sample size. This is perhaps the reason why the confidence intervals are wide; a larger sample should narrow the confidence intervals. Hence, we aim to increase this number in future research. Second, this paper has employed a RES approach to sampling the C_{IS} network. RES is better than random node sampling because it produces a sub network that resembles the underlying network better than the latter method; however, there are other more complex network sampling algorithms that improve upon RES. Utilizing a sampling approach that preserves network structure such as the hybrid random node-edge sampling method (Leskovec and Faloutsos 2006) should reveal useful information on how structural aspects of the network like network position influence new tie formation. Third, we defined collaboration as co-authorship in the Basket of 8 journals. However, some instances of collaboration do not result in publication in the Basket of 8 or at all. Comparing such collaboration to collaboration trends in the Basket of 8 may further isolate the unique aspects of collaboration in our flagship journals. Fourth, this study mainly examined the effects of node attributes in determining the probability of a tie between any two nodes. Incorporating edge attributes such as edge weight, method of research, and year of collaboration might also add extra insight on how authors choose to collaborate in the IS network. Finally, other factors such as researcher country of origin and race, complementary skill sets, similar interests, and co-location in socio-economic hubs, might also affect collaboration in the research network. Future work could add these variables to the model for validation.

Conclusion

In this paper, we examined the historical C_{IS} network by modeling it as a graph, with authors as vertices of the graph and edges representing co-authorship ties. We found that C_{IS} is scale-free and has grown in connectedness since 2012. We created a model of collaboration and found preliminary evidence that collaboration in IS demonstrates gender and geographical homophily; conversely we found evidence of strong diversity in temporal homophily. However, the results must be considered in the context of the

[5] http://dblp.uni-trier.de/

[6] http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1365-2575/homepage/special_issues.htm.

study's limitations, particularly that we conducted the analyses on only 110 edges out of a possible 10,000 edges, and we can gain more confidence in our findings as we collect more data. Lastly, we found that Kalle Lyytinen is the current center of C_{IS}. We defined the Lyytinen number and suggested how it can be used in evaluating a researcher's potential and/or past impact on the field.

Appendix

Data Collection

The first step was to retrieve all the tables of contents for all issues produced by the journals since April 1977 – the date of the first publication in MISQ. To accomplish this, we wrote scripts that employed the UNIX scripts *wget* and *curl* to send HTTP GET requests to these journals. The journal servers honored the requests and returned files containing the tables of contents of all journal issues published between 1977 and November 2015. We placed these files into eight different directories in preparation for parsing.

Parsing each HTML table of contents required the use of Java, regular expressions, and the Jsoup library. The goal of parsing the files was to retrieve the metadata for every published paper. We wrote a master parser that we customized for each different journal, because the HTML document tree structure varied across journals. The output of the parser was a map of publication titles and their respective co-authors. We wrote another Java program that created edges for each co-author tie. Each paper with at least two authors generated $n!/2$ edges, where n is the number of authors; for example a paper with 3 authors A, B and C generated 3!/2 links – A -> B, A -> C and B -> C. We did not incorporate directionality in constructing the network.

Lastly, we used the edge lists from the last step to create 1) the individual collaboration sub networks and 2) the combined network using the *R* programming language. All subsequent calculations on the network were done in *R* (Ihaka and Gentleman 1996). Figure 1 shows the process flow.

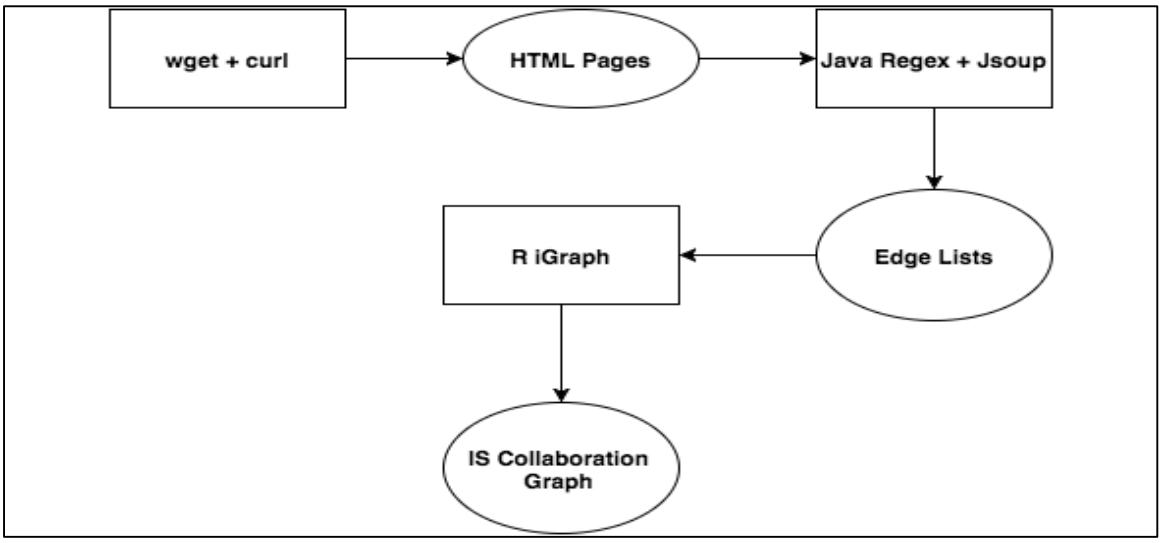

Figure 1: Method Process Flow

Technical Problems

There are various technical problems to be grappled with when collecting and parsing the data. First, two of the journal websites were not amenable to automated HTTP GET requests, possibly in fear of denial of service attacks. We had to devise workarounds in order to bypass their prevention mechanisms – particularly through wrapping our requests with *curl* rather than *wget* for one of the websites, and by reducing the average speed of our requests for the other website. Second, certain journals initially collected just the initials of the authors in lieu of their first names. There is no automated solution to this

problem; hence we had to manually search for the first names of authors in those journals. Last, in seeking demographic information for the modeling phase (described in the ergm sub-section), certain names are very common; examples are "Stephen Smith", "Susan Brown" and "Rui Chen." In such cases, specific collaboration information helped locate the correct individual for our analysis.

References

Akaike, H. 1974. "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on* (19:6), pp. 716–723.

Barabási, B. A.-L., and Bonabeau, E. 2003. "Scale-free," *Scientific American*.

Bonacich, P. 2007. "Some unique properties of eigenvector centrality," *Social networks* (29:4), pp. 555–564.

Coder, L., Rosenbloom, J. L., Ash, R. A., and Dupont, B. R. 2009. "Economic and business dimensionsIncreasing gender diversity in the IT work force," *Communications of the ACM* (52:5), p. 25.

Cuellar, M. J., Vidgen, R., Takeda, H., and Truex, D. 2016. "Ideational influence, connectedness, and venue representation: Making an assessment of scholarly capital," *Journal of the Association for Information Systems* (17:1), p. 1.

Currarini, S., Jackson, M. O., and Pin, P. 2009. "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica* (77:4), pp. 1003–1045.

De Castro, R., and Grossman, J. W. 1999. "Famous trails to Paul Erdos," *The Mathematical Intelligencer* (21:3), pp. 51–53.

Egorov, G., Polborn, M., and Welcome, C. A. V. 2010. "An Informational Theory of Homophily," (available at http://www.econ.upf.edu/docs/seminars/egorov.pdf).

Feld, S. L. 1982. "Social structural determinants of similarity among associates," *American Sociological Review*, pp. 797–801.

Gallivan, M., and Ahuja, M. 2015. "Co-authorship, Homophily, and Scholarly Influence in Information Systems Research," *Journal of the Association for Information Systems* (16:12), p. 980.

Gregor, S. 2006. "The nature of theory in information systems," *MIS quarterly*, pp. 611–642.

Grossman, J. W. 2002. "The Evolution of the Mathematical Research Collaboration Graph," *Congressus Numerantium* (158).

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. 2008. "statnet: Software tools for the representation, visualization, analysis and simulation of network data," *Journal of statistical software* (24:1), p. 1548.

Harris, J. K. 2013. *An introduction to exponential random graph modeling* (Vol. 173), Sage Publications.

Ihaka, R., and Gentleman, R. 1996. "R: a language for data analysis and graphics," *Journal of computational and graphical statistics* (5:3), pp. 299–314.

Joshi, A., and Roh, H. 2009. "The role of context in work team diversity research: A meta-analytic review," *Academy of Management Journal* (52:3), pp. 599–627.

Lazarsfeld, P. F., Merton, R. K., and others. 1954. "Friendship as a social process: A substantive and methodological analysis," *Freedom and control in modern society* (18:1), pp. 18–66.

Leskovec, J., and Faloutsos, C. 2006. "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 631–636.

Li, S. S., and Karahanna, E. 2015. "Online Recommendation Systems in a B2C E-Commerce Context: A Review and Future Directions," *Journal of the Association for Information Systems* (16:2), p. 72.

Lowry, P. B., Moody, G., Gaskin, J., Galletta, D. F., Humphreys, S., Barlow, J. B., and Wilson, D. 2013. "Evaluating journal quality and the association for information systems (AIS) senior scholars' journal basket via bibliometric measures: Do expert journal assessments add value?," *MIS Quarterly*, pp. 993–1012.

Mason, R. O., McKenney, J. L., and Copeland, D. G. 1997. "Developing an historical tradition in MIS research," *MIS quarterly*, pp. 257–278.

McPherson, M., Smith-Lovin, L., and Cook, J. M. 2001. "Birds of a feather: Homophily in social networks," *Annual review of sociology*, pp. 415–444.

Metcalfe, B. 1995. "Metcalfe's law: A network becomes more valuable as it reaches more users," *Infoworld* (17:40), pp. 53–54.

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. S. 1987. "Rediscovering the social group: A self-categorization theory," Basil Blackwell.

Wasserman, S., and Pattison, P. 1996. "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p," *Psychometrika* (61:3), pp. 401–425.

Wyatt, D., Choudhury, T., and Bilmes, J. A. 2008. "Learning Hidden Curved Exponential Family Models to Infer Face-to-Face Interaction Networks from Situated Speech Data.," in *AAAI*, pp. 732–738.

Xu, J., Chau, M., and Tan, B. C. 2014. "The development of social capital in the collaboration network of information systems scholars," *Journal of the Association for Information Systems* (15:12), p. 835.

Yuan, Y. C., and Gay, G. 2006. "Homophily of Network Ties and Bonding and Bridging Social Capital in Computer-Mediated Distributed Teams," *Journal of Computer-Mediated Communication* (11:4), pp. 1062–1084.

Zhai, L., Li, X., Yan, X., and Fan, W. 2014. "Evolutionary analysis of collaboration networks in the field of information systems," *Scientometrics* (101:3), pp. 1657–1677.

Zhang, P. 2015. "The IS History Initiative: Looking Forward by Looking Back," *Communications of the Association for Information Systems* (36:1), p. 24.