A Bibliometric and Network Analysis of the field of Computational Linguistics

Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson and Pradeep Muthukrishnan

$\{ radev, mtjoseph, gibsonb, mpradeep\}$@umich.edu

April 22, 2009

University of Michigan

Ann Arbor, MI 48109-1092

Abstract

The ACL Anthology is a large collection of research papers in computational linguistics. Citation data was obtained using text extraction from a collection of PDF files with significant manual post-processing performed to clean up the results. Manual annotation of the references was then performed to complete the citation network. We analyzed the networks of paper citations, author citations, and author collaborations in an attempt to identify the most central papers and authors. Also, we propose an improved method for comparing different measures of impact based on correlation. The analysis includes general network statistics, PageRank, metrics across publication years and venues, impact factor and h-index, as well as other measures.

1 Introduction

A typical outcome of a research project is a publication in a journal, conference, or other venue. Scientific papers cite each other and thus the ensemble of papers in a given field of research forms a directed network.

Analyzing the network of citations, we may be able to find interesting correlations that give us a new perspective on the importance of certain papers, their authors, the ideas presented in them, and the papers to which these important papers are connected.

In this paper we investigate the corpus of papers published by the Association for Computational Linguistics (ACL) by creating citation and collaboration networks and analyzing them using a variety of statistical measures.

With the help of these networks, we have been able to identify the most central papers and the most central authors according to different measures. We also disclose the identity of the Kevin Bacon of the AAN, i.e, the most central author in the collaboration network (Tjaden, 1996). We also analyze the correlation between the different ranking measures to identify if there is any single aspect that all the ranking measures value high. We were also able to analyze and observe interesting patterns about the overall impact of different venues in the field of computational linguistics over time. We also studied the effect of self citations, that is, an author citing his/her previous work, on the ranking of authors based on different measures.

In the next section we review previous research on citation and collaboration networks. In section 3, we describe the ACL Anthology. In section 4 we describe the measures used in the analysis. In sections 5 and 6 we present the networks created and the findings of our analysis. Finally, we discuss our conclusions and future work to be performed.

2 Related Work

Recently, there have been many papers (Albert and Barabasi, 2001; Newman, 2003; Dorogovtsev and Mendes, 2002; da F. Costa et al., 2007) on the analysis of real-world networks. With the ability to accumulate large amounts of information automatically, analysis of large-scale networks has become much easier than in the past. Some of the networks that have been studied are the World Wide Web (WWW), the Internet, citation networks, movie actor collaboration network, the web of human sexual contacts as well as power and neural networks. (Newman, 2003) showed that real-world networks are very different from random graphs (Erdös and Rényi, 1961), using an empirical analysis of the network properties. Network properties like average shortest path length, clustering coefficients, degree distribution and spectral properties clearly help distinguish between a random network and a real-world network. Based on these analyses, different models of evolution of the citation networks have been proposed.

Numerous papers have been published regarding collaboration networks in scientific journals, resulting in a number of important conclusions. In (Elmacioglu and Lee, 2005), it was shown that the DBLP network resembles a small-world network due to the presence of a high number of clusters with a small average distance between any two authors. This average distance is compared to (Milgram, 1967)'s "six degrees of separation" experiments, resulting in the DBLP measure of average distance between two authors stabilizing at approximately six. Similarly, in (Nascimento et al., 2003), the current (as of 2002) largest connected component of the SIGMOD network is identified as a small-world network, with a clustering coefficient of 0.69 and an average path length of 5.65.

Citation networks have also been the focus of recent research, with added concentration on the proceedings of major international conferences, and not just on leading journals in the scientific fields. In (Rahm and Thor, 2005), the contents over 10 years of the SIGMOD and VLDB proceedings along with the TODS, VLDB Journal, and SIGMOD Record were combined and analyzed. Statistics were provided for total and average number of citations per year. Though there are concerns as to its validity (Editorial, 2005), impact factor was also considered for the journal publications. Lastly, the most cited papers, authors, author institutions and their countries were found. In the end, they determined that the conference proceedings achieved a higher impact factor than journal articles, thus legitimizing their importance.

Citation networks other than paper citation networks (for example, the citation networks of legal court cases or patents) have also been studied. Patents cite other patents for a variety of reasons, mostly to establish their novelty over previous work. In legal citation networks, legal opinions cite other cases to establish precedent. One such network, the network of opinions of the United States Supreme Court was analyzed extensively in (Leicht et al., 2007). Leicht et al, have proposed a mixture model of citation patterns to discover community structure in citation networks. The hypothesis they put forth is that there exists a community structure in citation networks, each distinctly identifiable by its citation pattern. They use the Expectation Maximization algorithm (EM) to fit the mixture model they developed to the observed citation data. Then they apply Kleinberg's eigenvector centrality measure (Kleinberg, 1999) to the citation network to observe the top authority scores over time and reveal interesting facts about the evolution of the network. In particular, the plot of the average age of the top k authorities over time shows that the average age increases with time but with sudden drops. This shows that the top authorities remain the same for a substantial period of time but are swiftly replaced by younger leaders.

Another interesting aspect of citation networks and information diffusion was addressed in (Shi et al., 2008). Shi et al. addressed the question of what features are predictive of the popularity a paper would obtain in the citation graph. They found that papers which cite other recent papers in the same community garner a lot of citations over time. On the contrary, the most influential papers interdisciplinary and come out of ideas fused across communities. They also observed that, the citation structure in Computer Science depends on the area of research and the time period.

Interesting work has also been done regarding the correlation between the availability of articles and the number of citations those articles receive (Lawrence, 2001).

3 The ACL Anthology

The Association for Computational Linguistics (ACL) is an international professional society dedicated to the advancement in Natural Language Processing and Computational Linguistics research. The ACL Anthology, moderated by Steven Bird and Min Yen Kan, is a collection of papers from a journal published by ACL - $Computational\ Linguistics$ - as well as all proceedings from ACL sponsored conferences and workshops (`http://www.aclweb.org/anthology-new`) (Bird et al., 2008). It is from these papers that the ACL Anthology Network (AAN) was constructed (Joseph and Radev, 2007).

Table 1 includes a listing of the different conferences and the meeting years analyzed in this work, as well as the years for the ACL journal, *Computational Linguistics.* This represents the contents and standing of the ACL Anthology in February 2007. Since then, the proceedings of SIGDAT (Special Interest Group for linguistic data and corpus-based approaches to NLP) of the ACL have been extracted from the workshop heading and categorized separately.

Individual workshop listings have not been included in Table 1 due to space constraints. The assigned prefixes intended to represent each forum of publication are also included. These will be referenced in numerous tables within the paper and should make it easier to find the original conference or paper. For example, the proceedings of the European Chapter of the Association for Computational Linguistics conference have been assigned "E" as a prefix. So the ACL ID E02-1005 is a paper presented in 2002 at the EACL conference and assigned number 1005. It must be noted that not every year has been completed, as articles from HLT-02 are still absent.

In total, the ACL Anthology contains 11,749 unique papers from these various sources. Certain texts that did not include citations were not included such as Table of Contents, Front Matter, Author Index, Book Review, etc.

The AAN website (`http://belobog.si.umich.edu/clair/anthology/index.cgi`) displays all the statis-

Name	Prefix	Meeting Years
ACL	P	79-83, 84 w/COLING, 85-96, 97 w/EACL, 98 w/COLING, 99-05, 06 w/COLING
COLING	C	65, 67, 69, 73, 80, 82, 84 w/ACL, 86, 88, 90, 92, 94, 96, 98 w/ACL, 00, 02, 04 , 06 w/ACL, 07
EACL	E	83, 85, 87, 89, 91, 93, 95, 97 w/ACL, 99, 03, 06
NAACL	N	00 w/ANLP, 01, 03 w/HLT, 04 w/HLT, 06 w/HLT, 07
ANLP	A	83, 88, 92, 94, 97, 00 w/NAACL
SIGDAT (EMNLP & VLC)	D	93, 95-00, 02-04, 05 w/HLT, 06
TINLAP	T	75, 78, 87
Tipster	X	93, 96, 98
HLT	H	86, 89-94, 01, 03 w/NAACL, 04 w/NAACL, 05 w/EMNLP, 06 w/NAACL
MUC	M	91-93, 95
IJCNLP	I	05
Workshops	W	90-91, 93-07
Computational Linguistics	J	74-05

Table 1: *ACL Conference Proceedings. This includes the years for which analysis was performed.*

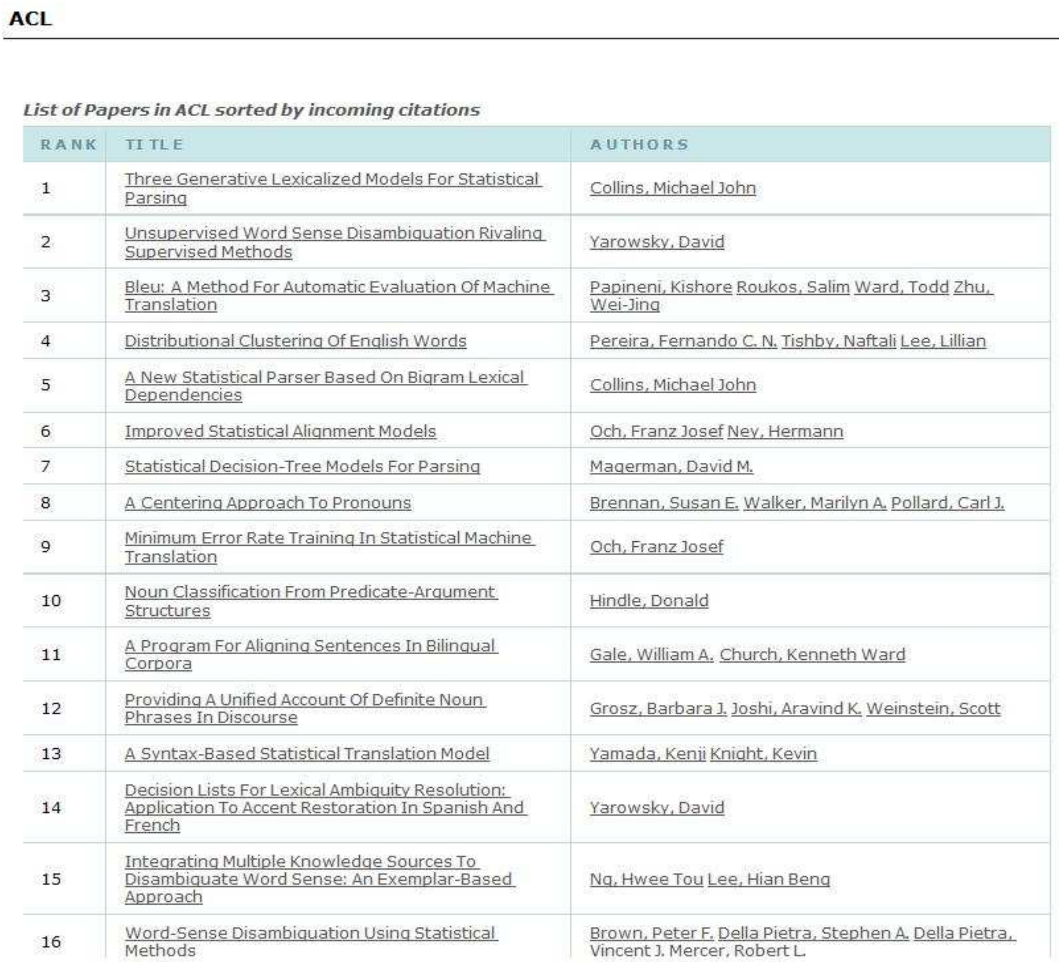
tics computed in this paper, the different rankings, and also includes features to select papers by conference as shown in Table 1 and search by author name, paper id, paper title, etc. A snapshot of the search feature is shown in Table 2.

Each of the papers was processed using an OCR text extraction tool (http://www.pdfbox.org/) and the references from each paper were parsed and extracted. The OCR text extraction outputs all the references as a single block and we had to manually insert line breaks between references. These references were then manually matched to other papers in the ACL Anthology using an “n-best” (with $n = 5$) matching algorithm built into a CGI interface. A snapshot of this interface is shown in Figure 3. The matched references were then compiled to produce a citation network. References to papers outside of the ACL were recorded but not included in the network. The statistics of the anthology citation network in comparison to the total number of references in the 11,749 papers can be seen in Table 2.

Total Papers Processed	11,749
Total Citations	167,165
Citations to papers within the Anthology	44,138, (26.4%)
Citations to other papers	123,023 (73.6%)

Table 2: *General Statistics. A citation is considered to be inside the anthology if it points to another paper in the ACL Anthology Network*

This process was very time consuming due to the sheer amount of data available and all the data inconsistencies that were encountered. An estimated 1,100 hours were spent on the extraction of the citations alone. Around 60% of the time was spent on matching the reference text to the correct papers using a user interface, 30% of the time on formatting the text version of the papers so that we can extract the references individually, 8% of the time on cleaning



ACL

List of Papers in ACL sorted by incoming citations

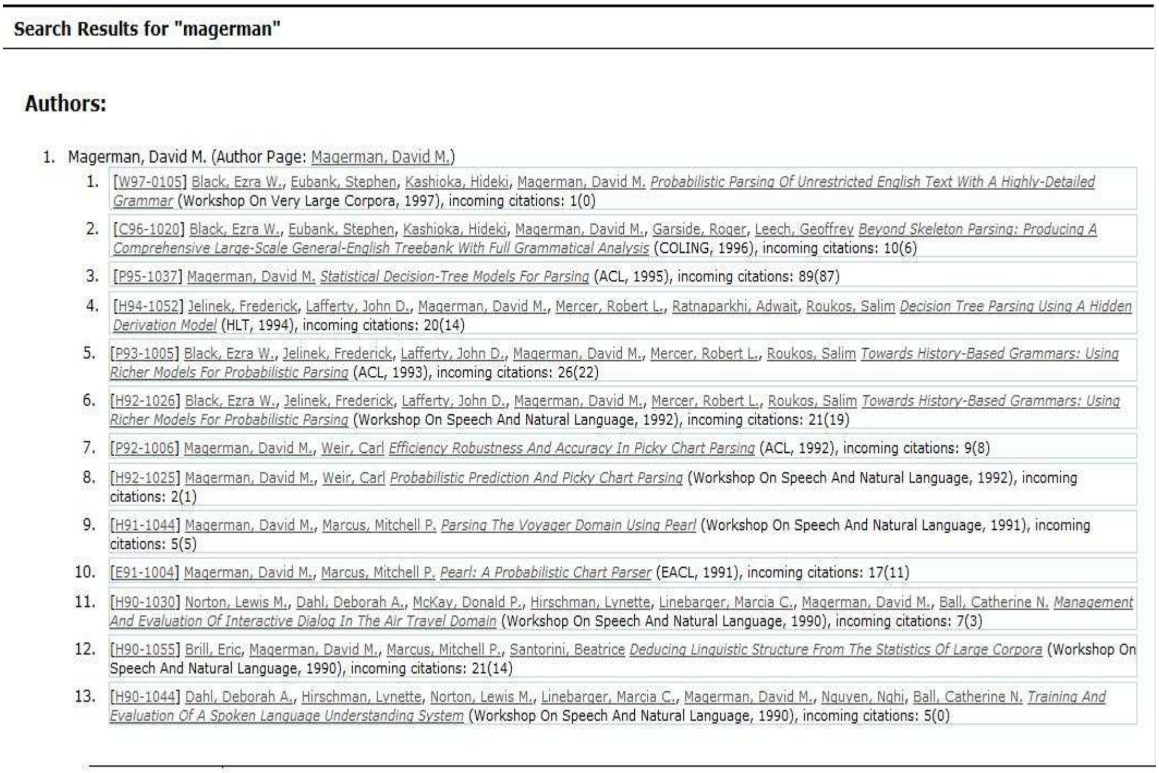
RANK	TITLE	AUTHORS
1	Three Generative Lexicalized Models For Statistical Parsing	Collins, Michael John
2	Unsupervised Word Sense Disambiguation Rivaling Supervised Methods	Yarowsky, David
3	Bleu: A Method For Automatic Evaluation Of Machine Translation	Papineni, Kishore Roukos, Salim Ward, Todd Zhu, Wei-Jing
4	Distributional Clustering Of English Words	Pereira, Fernando C. N. Tishby, Naftali Lee, Lillian
5	A New Statistical Parser Based On Bigram Lexical Dependencies	Collins, Michael John
6	Improved Statistical Alignment Models	Och, Franz Josef Ney, Hermann
7	Statistical Decision-Tree Models For Parsing	Magerman, David M.
8	A Centering Approach To Pronouns	Brennan, Susan E. Walker, Marilyn A. Pollard, Carl J.
9	Minimum Error Rate Training In Statistical Machine Translation	Och, Franz Josef
10	Noun Classification From Predicate-Argument Structures	Hindle, Donald
11	A Program For Aligning Sentences In Bilingual Corpora	Gale, William A. Church, Kenneth Ward
12	Providing A Unified Account Of Definite Noun Phrases In Discourse	Grosz, Barbara J. Joshi, Aravind K. Weinstein, Scott
13	A Syntax-Based Statistical Translation Model	Yamada, Kenji Knight, Kevin
14	Decision Lists For Lexical Ambiguity Resolution: Application To Accent Restoration In Spanish And French	Yarowsky, David
15	Integrating Multiple Knowledge Sources To Disambiguate Word Sense: An Exemplar-Based Approach	Ng, Hwee Tou Lee, Hian Beng
16	Word-Sense Disambiguation Using Statistical Methods	Brown, Peter F. Della Pietra, Stephen A. Della Pietra, Vincent J. Mercer, Robert L.

Figure 1: *Papers selected by a conference (ACL)*

up the data and correcting the citation data and 2% of the time in getting the different files in the right format and setting up the whole system.

In addition to the paper citation network, an author citation network and an author collaboration network were also created. The creation of these networks is described in detail in section 6. In attempting to build these author networks it was essential that we identify the correct authors for each paper. Aside from the casual misspelling of an author name, author names were sometimes missing from the publications. Often, a comma was lost or missing to indicate the appropriate order of first and last name resulting in *Klein Dan* (instead of *Dan Klein*). Also, authors sometimes use different versions of their name over the course of their publishing career (for instance, *Martha Stone* and *Martha Palmer*). An attempt to correct all such inconsistencies was made. The number of these issues was small in comparison to the vast number which were correct.

In section 4 we quickly go through the network analysis methods that will be used extensively in the later sections. Using these methods we analyze the connectedness, power law distributions in the paper citation network and the

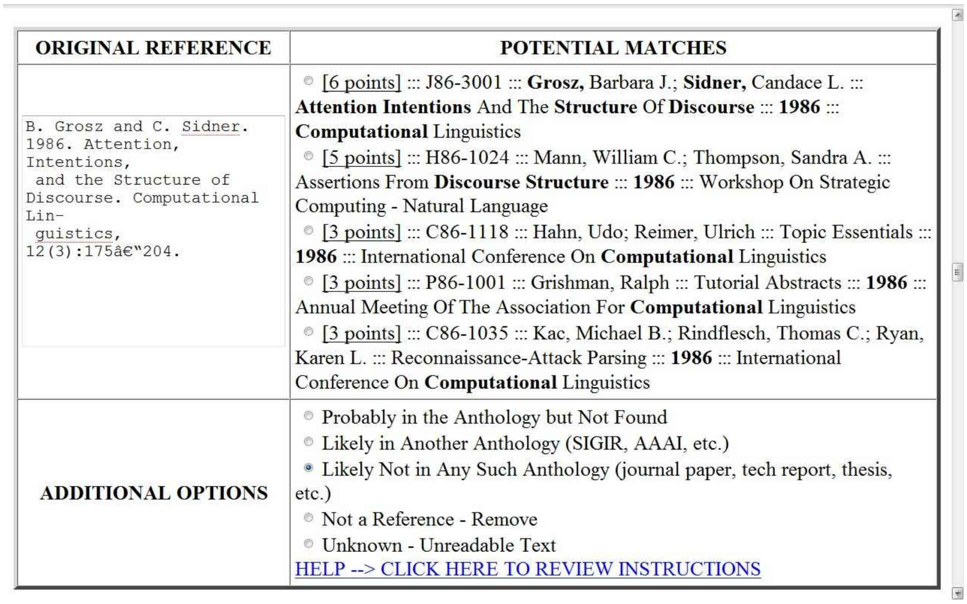


Search Results for "magerman"

Authors:

1. Magerman, David M. (Author Page: Magerman, David M.)
    1. [W97-0105] Black, Ezra W., Eubank, Stephen, Kashioka, Hideki, Magerman, David M. *Probabilistic Parsing Of Unrestricted English Text With A Highly-Detailed Grammar* (Workshop On Very Large Corpora, 1997), incoming citations: 1(0)
    2. [C96-1020] Black, Ezra W., Eubank, Stephen, Kashioka, Hideki, Magerman, David M., Garside, Roger, Leech, Geoffrey *Beyond Skeleton Parsing: Producing A Comprehensive Large-Scale General-English Treebank With Full Grammatical Analysis* (COLING, 1996), incoming citations: 10(6)
    3. [P95-1037] Magerman, David M. *Statistical Decision-Tree Models For Parsing* (ACL, 1995), incoming citations: 89(87)
    4. [H94-1052] Jelinek, Frederick, Lafferty, John D., Magerman, David M., Mercer, Robert L., Ratnaparkhi, Adwait, Roukos, Salim *Decision Tree Parsing Using A Hidden Derivation Model* (HLT, 1994), incoming citations: 20(14)
    5. [P93-1005] Black, Ezra W., Jelinek, Frederick, Lafferty, John D., Magerman, David M., Mercer, Robert L., Roukos, Salim *Towards History-Based Grammars: Using Richer Models For Probabilistic Parsing* (ACL, 1993), incoming citations: 26(22)
    6. [H92-1026] Black, Ezra W., Jelinek, Frederick, Lafferty, John D., Magerman, David M., Mercer, Robert L., Roukos, Salim *Towards History-Based Grammars: Using Richer Models For Probabilistic Parsing* (Workshop On Speech And Natural Language, 1992), incoming citations: 21(19)
    7. [P92-1006] Magerman, David M., Weir, Carl *Efficiency Robustness And Accuracy In Picky Chart Parsing* (ACL, 1992), incoming citations: 9(8)
    8. [H92-1025] Magerman, David M., Weir, Carl *Probabilistic Prediction And Picky Chart Parsing* (Workshop On Speech And Natural Language, 1992), incoming citations: 2(1)
    9. [H91-1044] Magerman, David M., Marcus, Mitchell P. *Parsing The Voyager Domain Using Pearl* (Workshop On Speech And Natural Language, 1991), incoming citations: 5(5)
    10. [E91-1004] Magerman, David M., Marcus, Mitchell P. *Pearl: A Probabilistic Chart Parser* (EACL, 1991), incoming citations: 17(11)
    11. [H90-1030] Norton, Lewis M., Dahl, Deborah A., McKay, Donald P., Hirschman, Lynette, Linebarger, Marcia C., Magerman, David M., Ball, Catherine N. *Management And Evaluation Of Interactive Dialog In The Air Travel Domain* (Workshop On Speech And Natural Language, 1990), incoming citations: 7(3)
    12. [H90-1055] Brill, Eric, Magerman, David M., Marcus, Mitchell P., Santorini, Beatrice *Deducing Linguistic Structure From The Statistics Of Large Corpora* (Workshop On Speech And Natural Language, 1990), incoming citations: 21(14)
    13. [H90-1044] Dahl, Deborah A., Hirschman, Lynette, Norton, Lewis M., Linebarger, Marcia C., Magerman, David M., Nguyen, Nghi, Ball, Catherine N. *Training And Evaluation Of A Spoken Language Understanding System* (Workshop On Speech And Natural Language, 1990), incoming citations: 5(0)

Figure 2: *Search results for "Magerman"*



ORIGINAL REFERENCE	POTENTIAL MATCHES
B. Grosz and C. Sidner. 1986. Attention, Intentions, and the Structure of Discourse. Computational Lin- guistics, 12(3):175â€"204.	○ [6 points] ::: J86-3001 ::: **Grosz,** Barbara J.; **Sidner,** Candace L. ::: **Attention Intentions** And The **Structure** Of **Discourse** ::: **1986** ::: **Computational** Linguistics ○ [5 points] ::: H86-1024 ::: Mann, William C.; Thompson, Sandra A. ::: Assertions From **Discourse Structure** ::: **1986** ::: Workshop On Strategic Computing - Natural Language ○ [3 points] ::: C86-1118 ::: Hahn, Udo; Reimer, Ulrich ::: Topic Essentials ::: **1986** ::: International Conference On **Computational** Linguistics ○ [3 points] ::: P86-1001 ::: Grishman, Ralph ::: Tutorial Abstracts ::: **1986** ::: Annual Meeting Of The Association For **Computational** Linguistics ○ [3 points] ::: C86-1035 ::: Kac, Michael B.; Rindflesch, Thomas C.; Ryan, Karen L. ::: Reconnaissance-Attack Parsing ::: **1986** ::: International Conference On **Computational** Linguistics
ADDITIONAL OPTIONS	○ Probably in the Anthology but Not Found ○ Likely in Another Anthology (SIGIR, AAAI, etc.) ● Likely Not in Any Such Anthology (journal paper, tech report, thesis, etc.) ○ Not a Reference - Remove ○ Unknown - Unreadable Text HELP --> CLICK HERE TO REVIEW INSTRUCTIONS

Figure 3: *Snapshot of the CGI interface used for matching references of new papers to existing papers. The annotators can choose from multiple options or indicate that the paper should not be added to AAN.*

author networks. In section 5 we analyze the paper citation network in an attempt to identify the papers with the most impact using different measures of centrality. In section 6 we analyze the author citation network and the author collaboration network. We look at different evaluation measures for ranking authors according to their impact by analyzing the author citation network. Also we look at the correlation between the centrality in the citation network and the collaboration network.

Before we begin looking at the analysis of the network, we describe some of the measures used in this analysis.

4 Network Analysis Methods

A network or a graph is usually represented as $G = (V, E)$, where V is the set of nodes and E is the set of edges. In case of directed networks, an edge is an ordered pair of vertices (u, v), that is, there exists an edge from u to v. On the contrary, in undirected networks, an edge is an unordered pair of vertices and does not have any direction. In the case of a weighted network, represented as $G = (V, E, w)$, each edge has an associated weight, typically given by the weight function $w : E- > R$. For example, the citation network is a directed, unweighted network while the collaboration network between authors is a weighted, undirected network. Both the networks were analyzed using Clairlib version 1.04[1].

4.1 Diameter and Average Shortest Path

In a network the average smallest number of steps along edges between any two nodes is called the *average shortest path*.

We computed two versions of average shortest path. The first is the sum of the length of shortest paths of all reachable node pairs, divided by the number of reachable pairs.

$$d_1 = \frac{\sum_{i=1}^{n}(\sum_{j=1}^{n} L_{ij})}{N_{rp}} n \tag{1}$$

where L_{ij} is the length of the shortest path from node i to node j, N_{rp} is the number of reachable pairs of nodes. The second comes from (Ferrer i Cancho and Solé, 2001), and is calculated as:

$$d_2 = \frac{\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n} L_{ij}}{n_i}\right)}{N} \tag{2}$$

where L_{ij} is the length of the shortest path from node i to node j, n_i is the number of neighbors of node i, and N is the number of nodes in the network.

Additionally, we calculated the *harmonic mean geodesic distance* as defined in (Newman, 2003). This measure gives an average of the distances between nodes, with lower values having a larger impact than higher outliers. In a network that does not allow self-loops, as is the case for the networks studied here, it is calculated as:

$$H = \frac{n(n-1)/2}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{L_{ij}}} \tag{3}$$

[1] Clairlib is a perl library designed by the University of Michigan Computational Linguistics And Information Retrieval (CLAIR) group (http://www.clairlib.org) (Radev et al., 2007)

Another common measure is the *network diameter*. The diameter of a graph is defined as the length of the longest shortest path between any two vertices.

For calculating the network diameter and distance, only the largest connected component of the network was used.

4.2 Power Law

One of the ways to identify the characteristics of a power law network's degree distribution is to calculate its power law exponent (α). A value of α near 2 indicates a network with power law characteristics.

We use two methods to calculate power law exponents. The first (α_{LS}) is a measure of the slope of the cumulative log-log degree distribution using the fitting of least squares (York, 1966). The power law exponent a is calculated as

$$\alpha = \frac{n \sum (x * y) - (\sum x * \sum y)}{(n * \sum x^2) - (\sum x)^2} \tag{4}$$

The r^2 statistic tells how well the linear regression line fits the data. The higher the value of r^2, the less variability in the fit of the data to the linear regression line. It is calculated as

$$r^2 = \frac{s_{xy}}{\sqrt{(s_{xx} * s_{yy})}} \tag{5}$$

where x is the independent variable, y the dependent variable, n the number or observations, and

$$s_{xy} = \frac{(\sum (x * y)) - (\sum x * \sum y)}{n}, \tag{6}$$

$$s_{xx} = \frac{\sum x^2 - (\sum x)^2}{n}, \tag{7}$$

$$s_{yy} = \frac{\sum y^2 - (\sum y)^2}{n} \tag{8}$$

The second calculation of the power law exponent (α_N) is modeled after (Newman, 2005)'s fifth formula, which is sensitive to a cutoff parameter that determines how much of the "tail" to measure. Newman's power law exponent α is calculated as

$$\alpha_N = 1 + n \left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}} \right]^{-1} \tag{9}$$

where n is the number of nodes in the network, x_i for $i = 1...n$ are the measured values of x, and x_{min} is the minimum value of x.

Newman's error is an estimate of the expected statistical error, and is calculated as

$$\sigma = \frac{\alpha - 1}{\sqrt{n}} \tag{10}$$

For example, Newman's power law exponent for a network where α = 2.500 and σ = 0.002 would estimate to α = 2.500 ± 0.002.

4.3 Clustering Coefficient

Finally, clustering coefficients are used to determine whether a network can be labeled as a small-world network. Two calculations were used.

The first clustering coefficient (Watts-Strogatz) (C_{WS}) in (Watts and Strogatz, 1998), is computed as

$$C_{WS} = \frac{\sum_i C_i}{n} \tag{11}$$

where n is the number of nodes and

$$C_i = \frac{T_i}{R_i} \tag{12}$$

with T_i defined as the number of triangles, or completely connected triples, connected to node i and R_i defined as the number of triples, both completely and partially connected, centered on node i.

The second (Newman) clustering coefficient (C_N) in (Newman, 2003), is computed as

$$C = \frac{3 * \text{ number of triangles in the network}}{\text{number of connected triples of vertices}} \tag{13}$$

For instance, if paper A cites paper B and paper B cites paper C, this is a connected triple. If paper A also cites paper C, or C cites A, then this is a triangle. Determining the number of triangles relative to the number of connected triples gives a measure of a network's transitivity. A real-world network will generally have a much higher clustering coefficient than a random network of the same size.

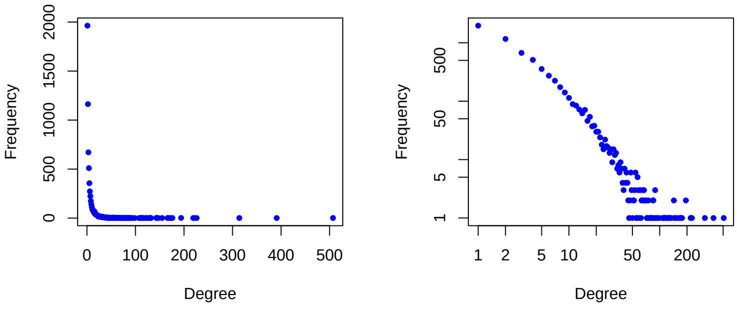
5 Paper Citation Network

The paper citation network includes all connections between ACL Anthology papers. It is a directed network with each node representing a paper labeled with an ACL ID number and the edges representing a citation within that paper to another paper represented by an ACL ID. The ACL ID number for each paper consists of a single letter denoting the venue and the year of publication, followed by the page number.

5.1 Paper Network - General Statistics

The network consists of 11,749 nodes, each representing a unique ACL ID number, and 44,138 directed edges. Of these nodes, 1,945 are completely disconnected with a degree of 0, leaving 9,764 connected nodes. The distribution of the in-degree, which is the number of citations a publication receives is shown in Figure 4. The size of the largest connected component is 9,594 with an average degree of 9.04, a diameter of 20, a clairlib average directed shortest path of 5.82, a Ferrer average directed shortest path of 5.11, a harmonic mean geodesic distance is 90.65, and an assortativity coefficient of 0.04. The paper citation network network contains 2,085 connected components. For this network C_{WS}=0.1879 and C_N=0.0804. A random network of the same size composed using the Erdos-Renyi model gives C_{WS}=0.0009 which is much lower (200x) than that of the AAN paper citation network, confirming that the AAN paper citation network is a small world network.

The power law values of the network are shown in Table 3. The value of α_N approaches 2, indicating a preference for edge attachment to a small number of high degree nodes.



(a) Standard scale

(b) Logarithmic scale

Figure 4: *Degree distribution of the paper citation network.*

Type	α_{LS}	r^2	α_N	σ
in-degree	2.52	0.97	2.03	0.02
out-degree	3.67	0.87	2.15	0.01
total degree	2.75	0.97	1.82	0.01

Table 3: *Paper Citation Network Power Law Measures*

5.2 Measures of Impact

In an effort to analyze the impact of the individual papers in the network, we looked at the total number of citations for each paper. The 20 most cited papers within the anthology are listed in Table 4.

Figure 5 shows the incoming citations by year from each year in the anthology regardless of venues. Recent years show a stronger occurence of reference than much older proceedings. This could be explained by the presence of higher numbers of papers in more recent years. The dominance of 1993 as a resource for citation does not fit well into the overall scheme until you consider that the two most cited papers in the anthology: *Building A Large Annotated Corpus Of English: The Penn Treebank* by Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (cited 507 times) and *The Mathematics Of Statistical Machine Translation: Parameter Estimation* by Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer (cited 391 times) were both published in *Computational Linguistics* in 1993.

5.3 Impact Factor

One popular measure of a venue's quality is its impact factor, one of the standard measures created by the Institute of Scientific Information (ISI).[2] Impact factor is calculated as follows:

[2] Institute of Scientific Information (ISI) is now acquired by Thomson Reuters Corporation and is now known as Thomson Scientific

ACL ID	Title	Authors	Number of Times Cited
J93-2004	Building A Large Annotated Corpus Of English: The Penn Treebank	Marcus, Mitchell P.; Marcinkiewicz, Mary Ann; Santorini, Beatrice	507
J93-2003	The Mathematics Of Statistical Machine Translation: Parameter Estimation	Brown, Peter F.; Della Pietra, Vincent J.; Della Pietra, Stephen A.; Mercer, Robert L.	391
J86-3001	Attention Intentions And The Structure Of Discourse	Grosz, Barbara J.; Sidner, Candace L.	314
A88-1019	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text	Church, Kenneth Ward	226
A00-2018	A Maximum-Entropy-Inspired Parser	Charniak, Eugene	221
J96-1002	A Maximum Entropy Approach To Natural Language Processing	Berger, Adam L.; Della Pietra, Vincent J.; Della Pietra, Stephen A.	219
P02-1040	Bleu: A Method For Automatic Evaluation Of Machine Translation	Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing	194
P97-1003	Three Generative Lexicalized Models For Statistical Parsing	Collins, Michael John	194
W96-0213	A Maximum Entropy Model For Part-Of-Speech Tagging	Ratnaparkhi, Adwait	176
J95-4004	Transformation-Based-Error-Driven Learning And Natural Language Processing: A Case Study In Part-Of-Speech Tagging	Brill, Eric	172
P95-1026	Unsupervised Word Sense Disambiguation Rivaling Supervised Methods	Yarowsky, David	166
J03-1002	A Systematic Comparison Of Various Statistical Alignment Models	Och, Franz Josef; Ney, Hermann	166
J02-3001	Automatic Labeling Of Semantic Roles	Gildea, Daniel; Jurafsky, Daniel	155
J90-2002	A Statistical Approach To Machine Translation	Brown, Peter F.; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Frederick; Lafferty, John D.; Mercer, Robert L.; Roossin, Paul S.	147
P03-1021	Minimum Error Rate Training In Statistical Machine Translation	Och, Franz Josef	143
J93-1003	Accurate Methods For The Statistics Of Surprise And Coincidence	Dunning, Ted E.	143
N03-1017	Statistical Phrase-Based Translation	Koehn, Philipp; Och, Franz Josef; Marcu, Daniel	143
J92-4003	Class-Based N-Gram Models Of Natural Language	Brown, Peter F.; DeSouza, Peter V.; Mercer, Robert L.; Watson, Thomas J.; Della Pietra, Vincent J.; Lai, Jennifer C.	132
J90-1003	Word Association Norms Mutual Information And Lexicography	Church, Kenneth Ward; Hanks, Patrick	131
J96-2004	Assessing Agreement On Classification Tasks: The Kappa Statistic	Carletta, Jean	130

Table 4: *20 Most Cited Papers in the Anthology*

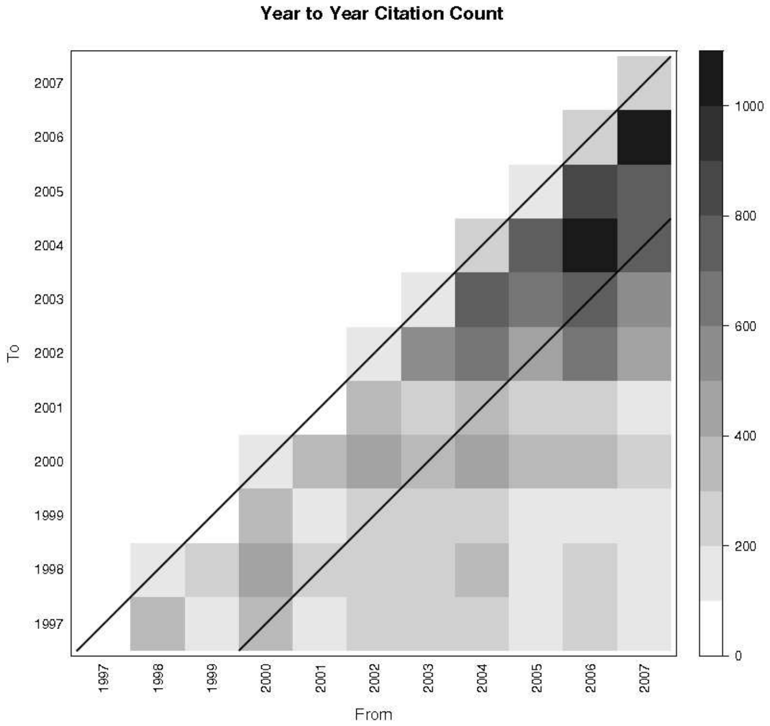


Figure 5: *Citation counts from one year to another (1997-2007). The area between the lines is a range of citations to the previous two years. Most papers cite recent papers.*

$$\frac{\text{Citations to Articles Published in Previous } k \text{ Years}}{\text{No. of Articles Published in Previous } k \text{ Years}} \tag{14}$$

For example, the impact factor over a two year period for a 2005 journal is equivalent to the number of citations included in that paper to publications in 2003 and 2004 divided by the total number of articles published in those two previous years (Amin and Mabe, 2000). This method may skew results in favor of popularity and not necessarily importance. Modifications and additional metrics have been proposed to account for this (Bollen et al., 2006), such as instead using a weighted PageRank or a combination of the two.

Impact factor was calculated for the ACL Anthology network based on a two year period using $k = 2$ in Equation 14. Figure 6 shows the results for each year where there is data in the AAN. In most of the years with lower impact (1989, 1995, 1999, 2001) there were fewer papers published than in neighboring years. Although 1985 had the same number of publications as the neighboring years, the number of citations from the publications published in 1985 to the previous two years' publications was less than average.

We also studied the impact of conferences and journals separately based on the number of citations they receive.

Table 5 shows the number of citations from papers in one type of publication to others, shown by year. (W=WS, J=CLJ, A=ANLP, N=NAACL, E=EACL, H=HLT, I=IJCNLP). For example, all ACL 2005 papers together included a

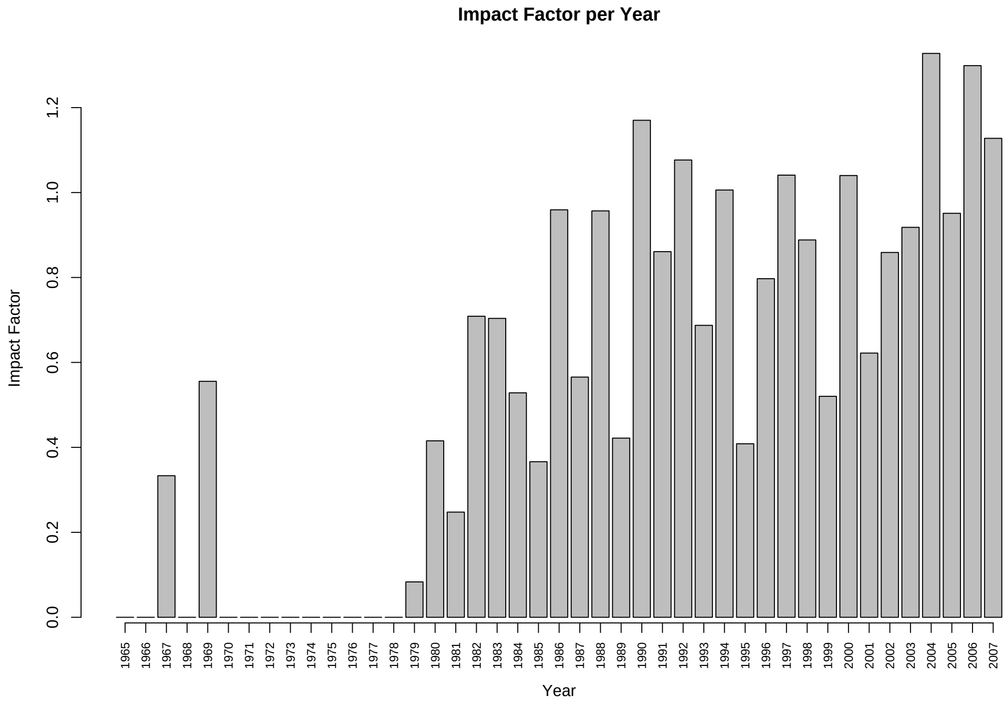


Figure 6: *Impact Factor per year from 1965 to 2007.*

total of 849 citations to other Anthology papers. Of these, 515 were to other conference papers, 191 were to workshop papers, and 143 were to (CL) journal papers.

This table shows that 75% of all citations in the journal to other Anthology papers go to conference and workshop papers and that 85% of all citations in ACL proceedings go to conference and workshop papers. In other words, on average a paper in ACL or $Computational\ Linguistics$ cites 4-5 times as many conference or workshop papers than journal papers.

Furthermore the percentage of citations from conference and workshop papers grows from year to year. In ACL 2007, 88% of its citations are from conference and workshop papers, compared with 78% in 2004.

This shows that conference and workshop papers are advancing the field and they are having more and more significant impact.

5.4 PageRank

The ClairLib library includes code to analyze the centrality of a network using the PageRank algorithm described in (Page et al., 1998). In calculating the ACL Anthology network centrality using PageRank, we find a general bias towards older papers. Older papers have had longer to accumulate new citations over time. It is not surprising then that the papers with the highest PageRank scores are slightly older. It should be noted that the PageRank scores are not accurate because of the lack of citations outside the AAN. Table 6 includes a listing of the 20 papers with the highest

W=WS, J=CLJ, A=ANLP, N=NAACL, E=EACL, H=HLT, I=IJCNLP

	W07	W06	W05	W04	W03	ACL07	ACL06	ACL05	ACL04	ACL03	N07	N06	N04	N03
A	51	77	58	97	79	31	70	45	26	33	15	20	18	29
C	106	208	112	179	144	93	222	69	54	71	36	45	29	29
E	70	77	28	40	29	49	55	16	16	12	25	11	6	6
H	74	118	13	47	35	61	119	11	15	9	47	51	15	15
J	225	330	241	414	180	163	287	143	144	104	90	103	71	73
M	6	10	1	8	20	5	4	2	4	7	2	1	4	3
ACL	566	760	318	540	337	456	663	281	220	162	293	219	152	117
W	660	772	407	560	319	313	465	191	135	117	165	162	104	75
X	0	1	0	0	0	0	1	1	0	0	0	0	0	0
N	158	188	111	125	37	134	147	90	50	13	98	89	48	13
TOTAL	1916	2541	1289	2010	1180	1305	2033	849	664	528	771	701	447	360
CONF	1031	1439	641	1036	681	829	1281	515	385	307	516	436	272	212
WS	660	772	407	560	319	313	465	191	135	117	165	162	104	75
JRNL	225	330	241	414	180	163	287	143	144	104	90	103	71	73
CONF(%)	0.54	0.57	0.5	0.52	0.58	0.64	0.63	0.61	0.58	0.58	0.67	0.62	0.61	0.59
WS(%)	0.34	0.3	0.32	0.28	0.27	0.24	0.23	0.22	0.2	0.22	0.21	0.23	0.23	0.21
JRNL(%)	0.12	0.13	0.19	0.21	0.15	0.12	0.14	0.17	0.22	0.2	0.12	0.15	0.16	0.2

	J05	J04	J03	I05	H05	E06	E03	C04	C02	A00
A	14	9	18	17	26	12	33	54	56	63
C	26	22	32	72	72	41	57	134	129	43
E	7	3	7	6	12	18	20	34	15	7
H	9	6	15	5	31	20	6	22	24	13
J	55	70	59	67	143	75	73	201	146	88
M	4	0	0	3	12	1	3	4	13	14
ACL	89	76	77	115	297	129	123	313	249	179
W	53	39	48	162	234	101	85	250	155	62
X	0	0	0	0	0	0	0	0	0	2
N	14	5	4	30	93	24	2	66	16	0
TOTAL	271	230	260	477	920	421	402	1078	803	471
CONF	163	121	153	248	543	245	244	627	502	321
WS	53	39	48	162	234	101	85	250	155	62
JRNL	55	70	59	67	143	75	73	201	146	88
CONF(%)	0.6	0.53	0.59	0.52	0.59	0.58	0.61	0.58	0.63	0.68
WS(%)	0.2	0.17	0.18	0.34	0.25	0.24	0.21	0.23	0.19	0.13
JRNL(%)	0.2	0.3	0.23	0.14	0.16	0.18	0.18	0.19	0.18	0.19

Table 5: *Inter-conference citation*

PageRank - rounded to the nearest ten-thousandth.

To address the fact that older papers have had a longer time period to accumulate incoming citations and hence will have higher PageRank values, we also calculated the PageRank per year for all of the papers in the ACL Anthology. To calculate this, we simply took the PageRank for each paper and divided by the number of years that had passed since that paper's publication. So, if a paper had been published in 2000, the PageRank would be divided by 8 (2008 minus 2000). Although this is not a widely studied statistic, we felt if may offer some further insight into the structure of the network. As one can see from the results in Table 7, this measure seems to favor the newer papers.

ACL ID	PageRank	Authors	Title
J93-2004	0.0062	Marcus, Mitchell P.; Marcinkiewicz, Mary Ann; Santorini, Beatrice	Building A Large Annotated Corpus Of English: The Penn Treebank
J93-2003	0.0050	Brown, Peter F.; Della Pietra, Vincent J.; Della Pietra, Stephen A.; Mercer, Robert L.	The Mathematics Of Statistical Machine Translation: Parameter Estimation
J86-3001	0.0070	Grosz, Barbara J.; Sidner, Candace L.	Attention Intentions And The Structure Of Discourse
J96-1002	0.0012	Berger, Adam L., Della Pietra, Vincent J., Della Pietra, Stephen A.,	A Maximum Entropy Approach To Natural Language Processing
A00-2018	0.0012	Charniak, Eugene	A Maximum-Entropy-Inspired Parser
P97-1003	0.0010	Collins, Michael John	Three Generative Lexicalized Models For Statistical Parsing
P02-1040	0.0010	Papineni, Kishore, Roukos, Salim, Ward, Todd, Zhu, Wei-Jing	Bleu: A Method For Automatic Evaluation Of Machine Translation
J95-4004	0.0009	Brill, Eric	Transformation-Based-Error-Driven Learning And Natural Language Processing: A Case Study In Part-Of-Speech Tagging
P95-1026	0.0009	Yarowsky, David	Unsupervised Word Sense Disambiguation Rivaling Supervised Methods
W96-0213	0.0008	Ratnaparkhi, Adwait	A Maximum Entropy Model For Part-Of-Speech Tagging
J03-1002	0.0008	Och, Franz Josef; Ney, Hermann	A Systematic Comparison Of Various Statistical Alignment Models
J02-3001	0.0008	Gildea, Daniel; Jurafsky, Daniel	Automatic Labeling Of Semantic Roles
J93-1003	0.0007	Dunning, Ted E.	Accurate Methods For The Statistics Of Surprise And Coincidence
J90-2002	0.0007	Brown, Peter F., Cocke, John, Della Pietra, Stephen A., Della Pietra, Vincent J., Jelinek, Frederick, Lafferty, John D., Mercer, Robert L., Roossin, Paul S.	A Statistical Approach To Machine Translation
J92-4003	0.0007	Brown, Peter F., DeSouza, Peter V., Mercer, Robert L., Watson, Thomas J., Della Pietra, Vincent J., Lai, Jennifer C.	Class-Based N-Gram Models Of Natural Language
N03-1017	0.0007	Koehn, Philipp, Och, Franz Josef, Marcu, Daniel	Statistical Phrase-Based Translation
P03-1021	0.0007	Och, Franz Josef	Minimum Error Rate Training In Statistical Machine Translation
J90-1003	0.0007	Church, Kenneth Ward, Hanks, Patrick	Word Association Norms Mutual Information And Lexicography

Table 6: *Papers with the Highest PageRanks*

ACL ID	PPY	Authors	Title
J93-2004	0.00019	Marcus, Mitchell P.; Marcinkiewicz, Mary Ann; Santorini, Beatrice	Building A Large Annotated Corpus Of English: The Penn Treebank
J03-1002	0.00017	Och, Franz Josef; Ney, Hermann	A Systematic Comparison Of Various Statistical Alignment Models
P02-1040	0.00016	Papineni, Kishore; Roukos, Salim; Ward, Todd; Zhu, Wei-Jing	Bleu: A Method For Automatic Evaluation Of Machine Translation
N03-1017	0.00015	Koehn, Philipp, Och, Franz Josef, Marcu, Daniel	Statistical Phrase-Based Translation
A00-2018	0.00015	Charniak, Eugene	A Maximum-Entropy-Inspired Parser
P03-1021	0.00014	Och, Franz Josef	Minimum Error Rate Training In Statistical Machine Translation
J93-2003	0.00014	Brown, Peter F.; Della Pietra, Vincent J.; Della Pietra, Stephen A.; Mercer, Robert L.	The Mathematics Of Statistical Machine Translation: Parameter Estimation
J02-3001	0.00013	Gildea, Daniel; Jurafsky, Daniel	Automatic Labeling Of Semantic Roles
P05-1033	0.00012	Chiang, David	A Hierarchical Phrase-Based Model For Statistical Machine Translation
P05-1022	0.00011	Charniak, Eugene; Johnson, Mark	Coarse-To-Fine N-Best Parsing And MaxEnt Discriminative Reranking
D07-1096	0.00010	Nivre, Joakim; Hall, Johan; Kubler, Sandra; McDonald, Ryan; Nilsson, Jens; Riedel, Sebastian; Yuret, Deniz	The CoNLL 2007 Shared Task on Dependency Parsing
J96-1002	0.00010	Berger, Adam L., Della Pietra, Vincent J., Della Pietra, Stephen A.,	A Maximum Entropy Approach To Natural Language Processing
P97-1003	0.00009	Collins, Michael John	Three Generative Lexicalized Models For Statistical Parsing
P03-1054	0.00009	Klein, Dan; Manning, Christopher D.	Accurate Unlexicalized Parsing
P00-1056	0.00008	Och, Franz Josef; Ney, Hermann	Improved Statistical Alignment Models
W06-2920	0.00008	Buchholz, Sabine; Marsi, Erwin	CoNLL-X Shared Task On Multilingual Dependency Parsing
J04-4002	0.00008	Och, Franz Josef; Ney, Hermann	The Alignment Template Approach To Statistical Machine Translation
W05-0620	0.00008	Carreras, Xavier; Marquez, Lluis	Introduction To The CoNLL-2005 Shared Task: Semantic Role Labeling
P05-1012	0.00007	McDonald, Ryan; Crammer, Koby; Pereira, Fernando C. N.	Online Large-Margin Training Of Dependency Parsers
P02-1038	0.00007	Och, Franz Josef, Ney, Hermann	Discriminative Training And Maximum Entropy Models For Statistical Machine Translation

Table 7: *Papers with the Highest PageRank per Year (PPY)*

6 Author Networks

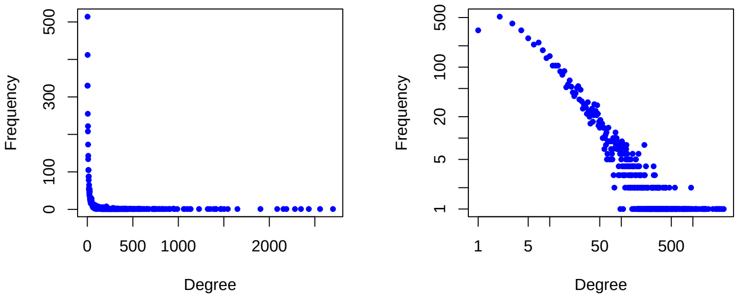
Using the paper network, and the metadata associated with each paper, we also created a network of author citations and a network of author collaborations. The following two sections describe in greater detail these two networks, as well as provide statistics and comparisons to other research.

6.1 Citation Network

The author citation network is derived from the paper network described previously where each node is a unique author and each edge is an occurrence of one author citing another author. For each paper, each author of that paper occurs as a node in the network. If one paper cites another paper, then all authors of the first paper cite all authors of the second paper. For example: if Andrea Setzer cites an earlier paper by James D. Pustejovsky, then the link "Setzer, Andrea → Pustejovsky, James D." will occur in the network. Self-citations are treated the same way. We have created two versions of the author citation network, one that includes self-citations and one that does not. Statistics from the network devoid of self-citations are shown in parentheses.

6.1.1 Citation Network - General Statistics

The author citation network consists of 9,421 (8,504) nodes and 158,497 (134,903) directed edges. The degree distribution can be seen in Figure 7. The size of the largest connected component is 7,672 (7,672) with a diameter of 10 (10), a clairlib avg. directed shortest path of 3.34 (3.34), a Ferrer avg. directed shortest path of 3.3 (3.3). The harmonic mean geodesic distance is 7.88 (7.88). Power law measures are given in Table 8. The power law measure given by the least squares method indicates a strong preference for new edges to attach to high degree nodes, while the Newman method gives a value showing a weaker preference.



(a) Standard scale

(b) Logarithmic scale

Figure 7: *Degree distribution of the author citation network.*

The clustering measures of this network are a C_{WS}=0.4687 (0.4584) and C_N=0.1474 (0.1374). In a random network of the same size, both C_{WS} and C_N would be 0.0017. The actual network values are significantly higher,

Type	α_{LS}	r^2	α_N	σ
in-degree	2.21 (2.21)	0.91 (0.91)	1.57 (1.57)	0.01 (0.01)
out-degree	2.57 (2.57)	0.85 (0.85)	1.55 (1.55)	0.00 (0.00)
total degree	2.28 (2.28)	0.89 (0.89)	1.46 (1.46)	0.00 (0.00)

Table 8: *Author Citation Network Power Law Measures. Refer to section 4 for an explanation of these measures.*

indicating a small world network.

6.1.2 Citation Network - Degree Statistics

In Table 9, we show the top 20 authors in both incoming and outgoing citations. Outgoing citations refer to the number of times an author cites other authors within the ACL Anthology. Incoming citations refer to the most cited authors within the ACL Anthology. It should be noted that the out-degree measure is expected to be proportional to the number of papers written by a specific author. In Table 10, the top 30 weighted edges are listed from the citation network. The weight represents the number of citations from one author to another. So, for instance, as one can see

Out-Degree (prolificness)		In-Degree (popularity)	
1367	Ney, Hermann	2699	Och, Franz Josef
1223	Tsujii, Jun'ichi	2557	Della Pietra, Vincent J.
1023	Marcu, Daniel	2433	Ney, Hermann
988	McKeown, Kathleen R.	2347	Mercer, Robert L.
848	Matsumoto, Yuji	2281	Della Pietra, Stephen A.
816	Hovy, Eduard H.	2187	Marcus, Mitchell P.
798	Collins, Michael John	2155	Church, Kenneth Ward
794	Grishman, Ralph	2086	Brown, Peter F.
760	Joshi, Aravind K.	1902	Collins, Michael John
758	Lapata, Mirella	1649	Yarowsky, David
723	Palmer, Martha Stone	1543	Charniak, Eugene
702	Koehn, Philipp	1502	Pereira, Fernando C. N.
657	Knight, Kevin	1469	Marcinkiewicz, Mary Ann
644	Miyao, Yusuke	1467	Grishman, Ralph
631	Carroll, John A.	1466	Santorini, Beatrice
625	Curran, James R.	1415	Joshi, Aravind K.
619	Ng, Hwee Tou	1408	Knight, Kevin
614	Wiebe, Janyce M.	1388	Brill, Eric
599	Johnson, Mark	1349	Marcu, Daniel
598	Och, Franz Josef	1323	Roukos, Salim

Table 9: *Author Citation Network Highest In- and Out-Degrees*

from the chart, Hermann Ney cites different works by Franz Josef Och 103 times. Individual papers may have multiple references to papers by the same author. It is common to cite your own research, which can be seen by the fact that 21 of the top 30 strongest edges in the graph are self-citations. This shows not only the prevalence of self-citation in research, but also points to a potential problem in networks of this type. The decision to include self-citations in a citation network will obviously skew the data in favor of authors who have written more papers and who use many self-citations.

An additional experiment performed was to calculate the log base 10 of the number of incoming citations for each paper for an author and to then sum these logs. This greatly reduced the skew of those authors with very large numbers of citations. The top 20 authors by this value are shown in Table 11.

6.1.3 Citation Network - h-index

In 2005, a new metric to calculate author prestige was proposed (Hirsch, 2005) called the h-index. ”A scientist has index h if h of their N papers have at least h citations each, and the other $(N - h)$ papers have no more than h citations each”. It is designed to highlight an author’s overall productivity, penalizing those authors who have only a few highly cited papers or many papers with low citing. There is some disagreement as to the usefulness of this metric as it appears to penalize younger authors and authors with fewer papers (Lehmann et al., 2006). Modifications to the calculation have been attempted to fix this deficiency (Sidiropoulos et al., 2006). Here, we continue to use the original method of computation as it continues to produce interesting results that match intuition (Hirsch, 2007).

One of the drawbacks of the h-index is that it can vary widely between different scientific disciplines, as well as between a broader discipline and one of its sub-disciplines. Using the author citation network, we attempt to look at how the h-index for a group of specialty publications, the ACL, compares to the h-index of those same researchers when calculated against their full publication history, approximated by their citations recorded in Google Scholar (GS).

We calculated the h-index for all authors in the AAN (h_{AAN}), but, due to space constraints, chose to only compare authors with an h-index of 9 or above, which amounted to 51 authors, against their GS h-index (h_{GS}). To find the h-index from GS, we used the Publish or Perish tool (Harzing, 2008). This tool queries GS to retrieve all publication data for each author entered. We queried the author names within all categories (science, humanities, etc.) due to the fact that many of the authors publishing in ACL venues also publish in venues devoted to other subjects (eg. Linguistics, Information Retrieval, Databases, Bioinformatics, Cognitive Science). Only articles and books were considered publications. Care was made to remove publications retrieved by name collisions or name misspellings, as well as records returned pertaining to patent submissions. The h_{GS} values were all recorded at the end of April, 2008 and reflect the current values at that time.

The resulting data can be found in Table 12. The average h_{AAN} for our sample is 10.63, with a high of 16 and a low of 9. The corresponding average for these authors for all h_{GS} is 27.08, with a high of 45 and a low of 11. The high values in GS are much higher than in the AAN, again due to the AAN being just a subset of the authors’ full

(168)	**Ney, Hermann → Ney, Hermann**
(122)	Ney, Hermann → Och, Franz Josef
(85)	**Tsujii, Jun’ichi → Tsujii, Jun’ichi**
(84)	**Grishman, Ralph → Grishman, Ralph**
(80)	**Joshi, Aravind K. → Joshi, Aravind K.**
(72)	Ney, Hermann → Della Pietra, Vincent J.
(71)	Ney, Hermann → Della Pietra, Stephen A.
(70)	Och, Franz Josef → Ney, Hermann
(69)	**Seneff, Stephanie → Seneff, Stephanie**
(68)	Ney, Hermann → Tillmann, Christoph
(64)	**Litman, Diane J. → Litman, Diane J.**
(64)	**Knight, Kevin → Knight, Kevin**
(62)	Ney, Hermann → Mercer, Robert L.
(62)	Ney, Hermann → Brown, Peter F.
(60)	Zens, Richard → Ney, Hermann
(60)	**Weischedel, Ralph M. → Weischedel, Ralph M.**
(60)	**Curran, James R. → Curran, James R.**
(59)	**Och, Franz Josef → Och, Franz Josef**
(59)	**Palmer, Martha Stone → Palmer, Martha Stone**
(57)	Zens, Richard → Och, Franz Josef
(57)	**Rambow, Owen → Rambow, Owen**
(57)	**McKeown, Kathleen R. → McKeown, Kathleen R.**
(56)	Curran, James R. → Clark, Stephen
(56)	**Johnson, Mark → Johnson, Mark**
(53)	**Clark, Stephen → Clark, Stephen**
(51)	**Schabes, Yves → Schabes, Yves**
(51)	**Wu, Dekai → Wu, Dekai**
(51)	**Bangalore, Srinivas → Bangalore, Srinivas**
(51)	**Marcu, Daniel → Marcu, Daniel**
(49)	**Hovy, Eduard H. → Hovy, Eduard H.**

Table 10: *Author Citation Network Highest Edge Weights. Bold values are self-citations.*

Log Sum	**Author**
34.63	Grishman, Ralph
33.42	Pereira, Fernando C. N.
31.43	Ney, Hermann
31.15	Church, Kenneth Ward
30.59	Joshi, Aravind K.
28.71	Johnson, Mark
28.44	Knight, Kevin
26.69	Hovy, Eduard H.
26.65	Manning, Christopher D.
26.60	McKeown, Kathleen R.
26.18	Och, Franz Josef
26.08	Marcu, Daniel
26.06	Yarowsky, David
25.80	Collins, Michael John
24.84	Charniak, Eugene
23.22	Brill, Eric
22.29	Mercer, Robert L.
21.97	Schabes, Yves
21.56	Moore, Robert C.
21.28	Palmer, Martha Stone

Table 11: *Author Citation Network - Incoming Citations Log Sums. Value is the sum of logs base 10 of incoming citations for each paper authored.*

publication history.

The Pearson correlation of the h_{GS} to the h_{AAN} is 0.51 for those authors with an h_{AAN} of 9 or above. The fairly low correlation shows that a high h_{GS} does not necessarily mean a high h_{AAN}. This is most likely due to the fact that some authors produce most of their highly cited work within the field covered by the ACL, while others produce most of their highly cited work outside of this field. For instance, Hermann Ney has published much in the speech community, leading to a much higher h_{GS} than h_{AAN}. The same is true of Fernando Pereira, publishing many papers in the machine learning community. To test the suspicion that authors with a much higher h_{GS} than h_{AAN} publish a significant amount outside of the AAN we did a regression of the AAN vs. GS h-index scores, shown in Figure 8. Author's more than 2 σ away from this line have an abnormal AAN to GS h-index ratio. The two authors who fall $\geq$ 2 σ above the line, Marti A. Hearst and Eduard H. Hovy, have many more highly cited papers outside of AAN than within AAN. Their h_{AAN}, using a subset of their papers, was significantly lower than their overall h-index. The author who falls below -2σ, Stephen Clark, has published all of his papers within AAN. The AAN index here is representative of the total h-index for the author. Another correlation tested was that of h_{AAN} against the author's incoming citation

Author	**AAN h-index**	**GS h-index**
Church, Kenneth Ward	16	38
Knight, Kevin	15	32
Grishman, Ralph	14	30
Joshi, Aravind K.	14	33
Ney, Hermann	14	45
Pereira, Fernando C. N.	14	45
Yarowsky, David	13	30
Collins, Michael John	12	24
Manning, Christopher D.	12	32
Marcu, Daniel	12	32
McKeown, Kathleen R.	12	39
Mercer, Robert L.	12	35
Och, Franz Josef	12	25
Schabes, Yves	12	25
Shieber, Stuart M.	12	34
Brill, Eric	11	23
Charniak, Eugene	11	37
Dagan, Ido	11	24
Johnson, Mark	11	20
Resnik, Philip	11	30
Carroll, John A.	10	28
Daelemans, Walter	10	30
Gale, William A.	10	27
Hirschman, Lynette	10	30
Hovy, Eduard H.	10	36
Jelinek, Frederick	10	34
Jurafsky, Daniel	10	33
Klein, Dan	10	18
Moore, Robert C.	10	22
Palmer, Martha Stone	10	25
Roukos, Salim	10	30
Weischedel, Ralph M.	10	25
Alshawi, Hiyan	9	19
Bangalore, Srinivas	9	16
Briscoe, Ted	9	29
Brown, Peter F.	9	22
Clark, Stephen	9	11
Della Pietra, Vincent J.	9	15
Gildea, Daniel	9	15
Hearst, Marti A.	9	39
Lee, Lillian	9	18
Marcus, Mitchell P.	9	20
Melamed, I. Dan	9	17
Mihalcea, Rada	9	25
Moens, Marc	9	22
Ng, Hwee Tou	9	17
Rambow, Owen	9	21
Riloff, Ellen	9	27
Tillmann, Christoph	9	15
Walker, Marilyn A.	9	32
Webber, Bonnie Lynn	9	30

Table 12: *Author Citation h-index - AAN vs. Google Scholar for AAN h-index* ≥ 9

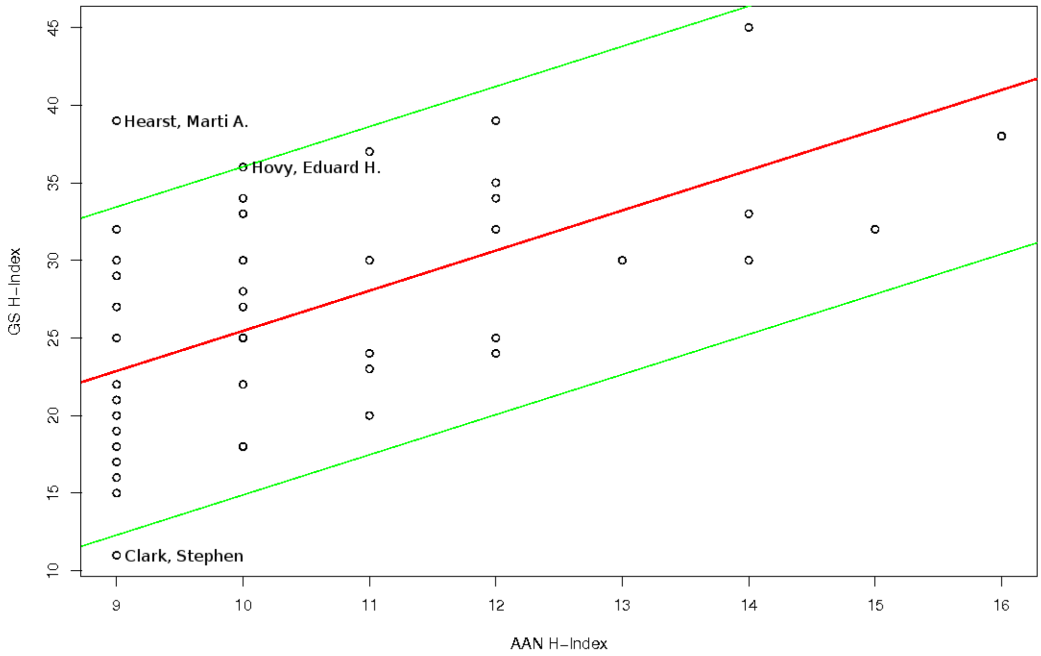


Figure 8: *AAN vs. GS h-index Regression. The thicker line represents the regression while the two thinner lines represent 2 σ from the regression.*

count within the AAN, again for authors with an AAN h-index of 9 or above. The Pearson correlation here was also low at 0.52. Figure 9 shows the calculated regression. All the authors above the line of regression have a small number of very highly cited papers. This is one argument against the h-index, that authors who might be considered central due to the importance of one or two of their papers are penalized. The last correlation we investigated was between an author's h_{AAN} and their PageRank in the author citation network. This is the weighted PageRank, where each citation from one author to another is counted as a weight between those authors. Again we used the same list of 51 top authors by h-index. The results can be seen in Figure 10. The correlation here is not very strong at all with a Pearson correlation coefficient of 0.33. All of the authors who appear as outliers are early pioneers who wrote very influential papers early on. Their papers have gained important links disproportionate to other authors in the list. In order to investigate where authors may be publishing papers outside of the AAN, we chose to look at one author and determine the venue of all of the papers that appear a Google Scholar search for that author which contribute to their h-index score. We chose the author 'Yarowsky, D' due to its rare spelling, making the search easier. The results are shown in Table 13. Out of 29 publications, only 15 are included in both AAN and GS, dramatically reducing the papers available for h_{AAN} as compared with h_{GS}.

In AAN?	Venue Type	Venue	Year	Title
Y	Conference	CoNLL	2003	*Unsupervised personal name disambiguation*
N	Journal	Natural Language Engineering	2003	*Combining Classifiers for word sense disambiguation*
Y	Conference	EMNLP	2002	*Modeling consensus: classifier combination for word sense disambiguation*
N	Report	Progress in Speech Synthesis	2002	*Evaluating sense disambiguation across diverse parameter spaces*
Y	Conference	HLT	2001	*Inducing multilingual text analysis tools via robust projection across aligned corpora*
N	Workshop	SENSEVAL2	2001	*The Johns Hopkins SENSEVAL2 system descriptions*
Y	Conference	NAACL	2001	*Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora*
Y	Conference	NAACL	2001	*Multipath translation lexicon induction via bridge languages*
Y	Conference	ACL	2000	*Minimally supervised morphological analysis by multimodal alignment*
Y	Conference	ACL	2000	*Rule writing or annotation: cost-efficient resource usage for base noun phrase chunking*
N	Journal	Computers and the Humanities	2000	*Hierarchical decision lists for word sense disambiguation*
N	Journal	Natural Language Engineering	2000	*Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation*
Y	Conference	ACL	1999	*Dynamic nonlocal language modeling via hierarchical topic-based adaptation*
N	Workshop	JHU Summer WS	1999	*Statistical machine translation*
Y	Conference	SIGDAT	1999	*Language independent named entity recognition combining morphological and contextual evidence*
N	Workshop	ACL SIGLEX	1997	*A perspective on word sense disambiguation methods and their evaluation*
N	Journal	Natural Language Engineering	1997	*Homograph disambiguation in text-to-speech synthesis*
Y	Conference	ACL	1995	*Unsupervised word sense disambiguation rivaling supervised methods*
N	Journal	Annals of Operations Research	1995	*Discrimination decisions for 100,000-dimensional spaces*
Y	Conference	ACL	1994	*Decision lists for lexical ambiguity resolution: application to accent restoration in Spanish and French*
N	Book	Natural Language Processing Using Very Large Corpora	1994	*A comparison of corpus-based techniques for restoring accents in spanish and french text*
Y	Conference	HLT	1993	*One sense per collocation*
N	Conference	AAAI	1992	*Work on statistical methods for word sense disambiguation*
Y	Conference	ACL	1992	*Estimating upper and lower bounds on the performance of word-sense disambiguation programs*
Y	Conference	COLING	1992	*Word-sense disambiguation using statistical models of Rogets categories trained on large corpora*
N	Journal	Computers and the Humanities	1992	*A method for disambiguating word senses in a large corpus*
N	Conference	ICSLP	1992	*A corpus-based synthesizer*
N	Conference	MT	1992	*Using bilingual materials to develop word sense disambiguation methods*
Y	Workshop	WS on Speech and Natural Language	1992	*One sense per discourse*

Table 13: *Author Citation h-index - Google Scholar Results Venues for Yarowsky, D*

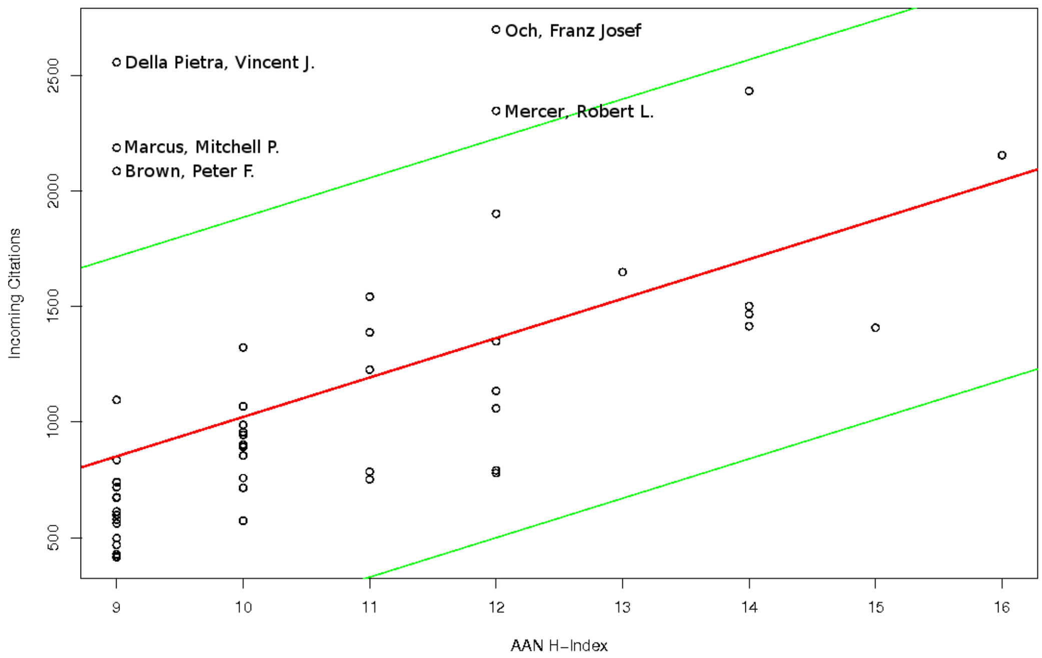


Figure 9: *AAN h-index vs. Incoming Citations. The thicker line represents the regression while the two thinner lines represent 2 σ from the regression.*

6.1.4 Citation Network - PageRank

We computed the PageRank centrality of the author citation network. For this measure, in order to avoid bias due to repeated citations, we analyzed two different networks, both an unweighted and a weighted citation network. The weighted network weights each edge with the number of repeated citations whereas the unweighted network treats all incidences of a citation from one author to another as a single occurrence.

The top weighted and unweighted PageRank results can be seen in Table 14. Values have been rounded to the nearest hundred-thousandth. Both weighted and unweighted networks still generally share the same central authors in the ACL Citation Network - 17 out of 20 authors show up in both lists.

6.1.5 Citation Network - Correlations between different measures of Impact

We performed several experiments in comparing the different measures of impact. Currently, there are various measures of impact proposed for citation networks. We computed various measures of impact in the author citation network such as h-index, total number of incoming citations, PageRank.

We computed the Pearson's rank correlation coefficient for each pair of the measures of the impact. Since we are more interested in finding out how the different metrics perform in ranking the authors at the top than ranking the bottom-ranked authors, we choose only the top k authors to compute the correlation coefficient. Since we are

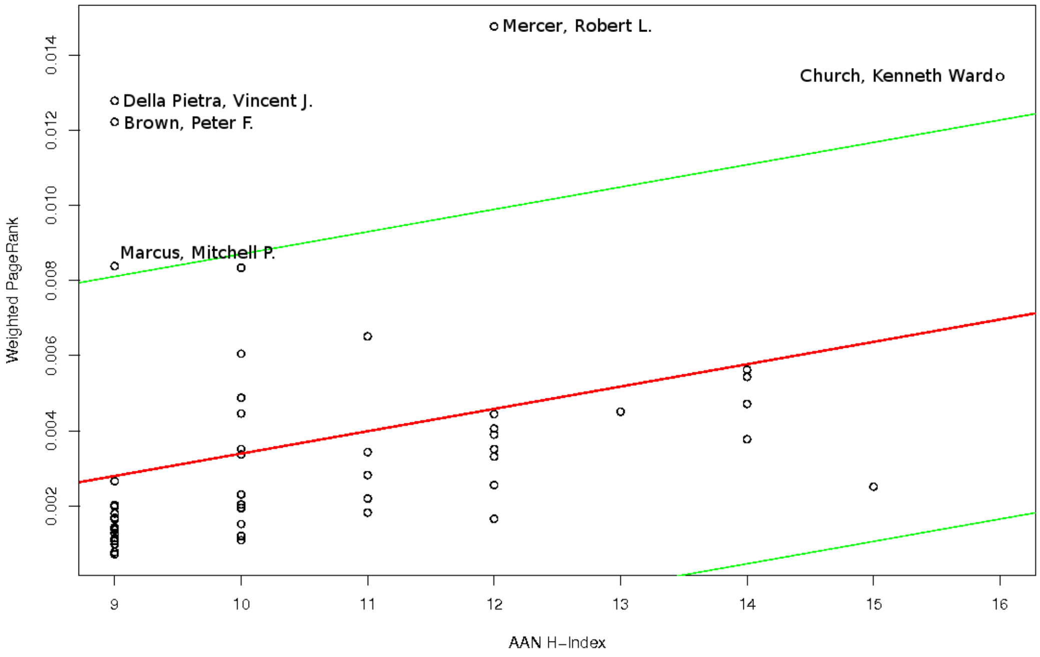


Figure 10: *AAN h-index vs. PageRank With Regression. The thicker line represents the regression while the two thinner lines represent 2 σ from the regression. PageRank is from the author citation network.*

choosing only top k authors, the coefficient depends on according to which metric we choose the top k authors. Therefore, we choose the top k authors according to both metrics and plot two curves for each pair of metrics that we are comparing. Suppose we are comparing two metrics A and B, then we choose the top i authors according to metric A for $1 \leq i \leq 200$ and plot the Pearson's rank correlation coefficient for each i. Similarly we choose the top i authors according to metrics B for $1 \leq i \leq 200$ and plot the correlation coefficient for each i. In the first plot, the hypothesis we are testing is does a high rank according to metric A imply a high rank according to metric B. In the second plot, the hypothesis we are testing is does a high rank according to metric B imply a high rank according to metric A. These correlation values are shown in Table 15. The metric along the rows is the metric according to which we sorted the authors. These correlation values are for the top 200 authors. The plots showing the correlation coefficient as number of authors chosen are increased are shown in 11.

For computing correlations, we had to decide if we should use the measures of impact including self citations or without including self citations. To help us decide, we computed the correlation between the different measures of impact including self citations and without including self citations. The Pearson Correlation Coefficient values were around 0.99 for the top 200 authors and 0.999 when we use all the authors. From these correlation coefficient values it is clear that it does not matter whether we include self citations or not for computing the correlations between the measures of impact.

Weighted		Unweighted	
Author	**PageRank**	**Author**	**PageRank**
Sampson, Geoffrey	0.01566	Church, Kenneth Ward	0.00606
Mercer, Robert L.	0.01333	Marcus, Mitchell P.	0.00595
Church, Kenneth Ward	0.01284	Della Pietra, Vincent J.	0.00582
Della Pietra, Vincent J.	0.01183	Mercer, Robert L.	0.00542
Brown, Peter F.	0.01147	Della Pietra, Stephen A.	0.00541
Della Pietra, Stephen A.	0.01084	Santorini, Beatrice	0.00519
Marcus, Mitchell P.	0.00774	Roukos, Salim	0.00508
Jelinek, Frederick	0.00714	Brown, Peter F.	0.00500
Brill, Eric	0.00616	Brill, Eric	0.00490
Weischedel, Ralph M.	0.00553	Collins, Michael John	0.00486
Grosz, Barbara J.	0.00547	Marcinkiewicz, Mary Ann	0.00477
Joshi, Aravind K.	0.00522	Grishman, Ralph	0.00476
Pereira, Fernando C. N.	0.00521	Pereira, Fernando C. N.	0.00465
Santorini, Beatrice	0.00492	Jelinek, Frederick	0.00464
Hindle, Donald	0.00481	Hindle, Donald	0.00441
Lafferty, John D.	0.00478	Weischedel, Ralph M.	0.00412
Grishman, Ralph	0.00447	Yarowsky, David	0.00409
Yarowsky, David	0.00446	Ratnaparkhi, Adwait	0.00391
Gale, William A.	0.00422	Ramshaw, Lance A.	0.00388
Schwartz, Richard M.	0.00414	Schwartz, Richard M.	0.00378

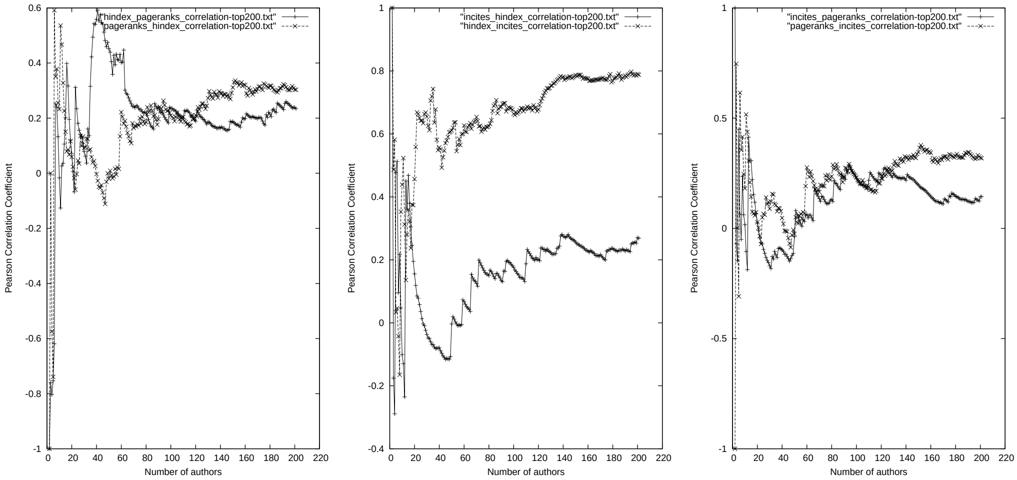
Table 14: *Author Citation Network PageRank. Weighted includes multiple citation from author A to author B, while unweighted removes multiple citations.*

The most interesting fact from these plots is that, when we sort the authors according to their h-index and then compute the correlation between h-index and Total Incoming Citations as a function of number of authors, we get a much higher correlation than when we sort the authors according to Total Incoming Citations and then compute the correlation. In the former method of computing correlation, the hypothesis we are testing is whether a high h-index imply a high number of Total Incoming Citations. Similarly, in the latter method, the hypothesis being tested is does a high number of Total Incoming Citations imply a high h-index. From the curves, it is clear that a high h-index implies a high number of Total Incoming Citations, whereas the converse is not true. This means that, the h-index is a better measure for ranking authors than the number of Total Incoming Citations.

To further analyse this, we plotted the number of incoming citations and h-index of every author as a scatter plot. The scatter plot is shown in Fig 12. It can be seen from this plot that there are many authors with the same number of incoming citations but different h-index. This is because of the fact that h-index is not a function of the total number of incoming citations but rather a function of the distribution of the incoming citations.

	h-index	Total citations	PageRank
h-index	1.0	0.79	0.24
Total citations	0.27	1.0	0.14
PageRank	0.30	0.32	1.0

Table 15: *Author Citation Network - Correlations between measures of impact*



(a) Correlation between h-index and PageRank

(b) Correlation between h-index and Incoming citations.

(c) Correlation between Incoming Citations and PageRank

Figure 11: *Correlations between measures of impact in the author citation network*

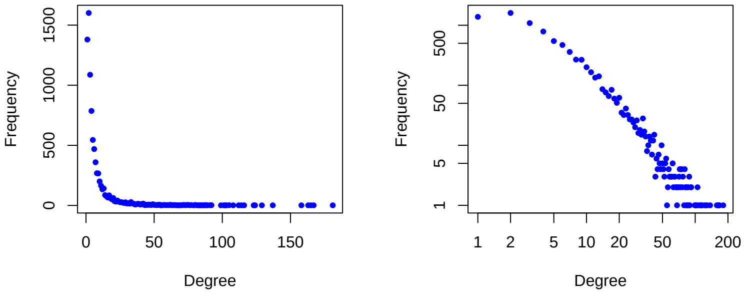
Figure 12: *Incoming Citations as a function of h-index*

6.2 Collaboration Network

The author collaboration network is based on the metadata of the ACL Anthology. Whenever one researcher co-authors (or collaborates on) a paper with another author, an edge between the two is recorded. For instance, the paper “Balancing Data-Driven And Rule-Based Approaches In The Context Of A Multimodal Conversational System” was authored by Srinivas Bangalore and Michael Johnston. This collaboration is recorded as the edge “Bangalore, Srinivas ↔ Johnston, Michael” in the network. Because of the way collaborations are inferred from authorship lists, it should be noted that this network is undirected.

6.2.1 Collaboration Network - General Statistics

The collaboration network consisted of 9421 nodes and 22,941 undirected edges. The degree distribution can be seen in Figure 13. The largest connected component is 7672 with a diameter of 20, a clairlib avg. directed shortest path of 5.86, a Ferrer avg. directed shortest path of 4.63 and a harmonic mean geodesic distance of 9.57. Power law exponent results can be found in Table 16. Note that because this network is undirected, only the total degree power law measure has been computed.



(a) Standard scale

(b) Logarithmic scale

Figure 13: *Degree distribution of the author collaboration network.*

Type	α_{LS}	r^2	α_N	σ
total degree	3.15	0.90	1.80	0.01

Table 16: *Author Collaboration Network Power Law Measures.*
Refer to section 4 for an explanation of these measures.

The power law values indicate that the network likely demonstrates characteristics of a power law relationship.

For this network, C_{WS}=0.6380 and C_N=0.3799 are much higher than in a random network of the same size, where C_{WS}=C_N=0.00025. The author collaboration network should be considered a small world network.

The results of other research are included in comparison to our findings for the ACL Anthology Network in Table 17. The power law values are similar, showing a similar propensity for papers with high numbers of citations to gain new citations. The clustering coefficient is different however with the DBLP appearing to be a much more well connected network.

6.2.2 Collaboration Network - Degree Statistics

In Table 18, we show the 20 authors with the most collaborations in the ACL Anthology Network and the number of collaboration they have been party to, where a collaboration is an event of one author publishing a paper with another

Archive	α_{LS}	C_N
DBLP (Elmacioglu and Lee, 2005)	3.68	0.63
ACL Anthology (this paper)	3.15	0.38

Table 17: *Author Collaboration Networks Comparison*

author. This equals the degree of the author's node in the collaboration network.

(181)	Tsujii, Jun'ichi	(114)	Wilks, Yorick
(167)	Hirschman, Lynette	(112)	Ingria, Robert J. P.
(165)	Weischedel, Ralph M.	(108)	McKeown, Kathleen R.
(163)	Schwartz, Richard M.	(105)	Hovy, Eduard H.
(158)	Isahara, Hitoshi	(105)	Matsumoto, Yuji
(137)	Grishman, Ralph	(103)	Waibel, Alex
(129)	Joshi, Aravind K.	(102)	Lavie, Alon
(124)	Ney, Hermann	(101)	Roukos, Salim
(123)	Rayner, Manny	(99)	Seneff, Stephanie
(116)	Palmer, Martha Stone	(92)	Zue, Victor W.

Table 18: *Author Collaboration Network - Most Collaborations*

Table 20 shows the top 34 weighted edges from the collaboration network. The edge weight n represents the number of times the two authors have collaborated together. Here the most central author is the author with the highest degree centrality. Using only unique edges for computing degree centrality, we find that Lynette Hirschman has the largest number of collaborators in AAN (87).

The Kevin Bacon of a network, however, is the person with the smallest shortest path to all other people in (giant connected component of) the network. In the AAN case, that honor belongs to Mitchell Marcus, whose average distance to the rest of the nodes in the giant component is only 2.62 steps.

1. Mitchell P. Marcus 2.620 2. Eduard H. Hovy 2.623 3. Owen Rambow 2.646 4. Ralph Grishman 2.675 5. Martha Stone Palmer 2.677 6. Kathl

Table 19: *Who is the Kevin Bacon of AAN?*

6.2.3 Collaboration Network - Shortest Paths

We also analyzed the shortest paths in the collaboration network. Since the network is unweighted and undirected, a simple Breadth First Search (BFS) was used to compute the shortest paths. The shortest path distance distribution is shown in Table 21. A value of -1 indicates that there is no path between a pair of authors. The distribution is plotted

Degree	Collaboration
(23)	Tsujii, Jun'ichi ↔ Miyao, Yusuke
(21)	Makhoul, John ↔ Schwartz, Richard M.
(19)	Uchimoto, Kiyotaka ↔ Isahara, Hitoshi
(18)	Zens, Richard ↔ Ney, Hermann
(17)	Murata, Masaki ↔ Isahara, Hitoshi
(17)	Joshi, Aravind K. ↔ Webber, Bonnie Lynn
(16)	Isahara, Hitoshi ↔ Ma, Qing
(15)	Rayner, Manny ↔ Hockey, Beth Ann
(15)	Zue, Victor W. ↔ Seneff, Stephanie
(15)	Och, Franz Josef ↔ Ney, Hermann
(14)	Pazienza, Maria Teresa ↔ Basili, Roberto
(14)	Bear, John ↔ Appelt, Douglas E.
(14)	Su, Jian ↔ Zhou, GuoDong
(14)	Curran, James R. ↔ Clark, Stephen
(14)	Lin, Chin Yew ↔ Hovy, Eduard H.
(14)	Grishman, Ralph ↔ Sterling, John
(13)	Wu, Dekai ↔ Carpuat, Marine
(13)	Phillips, Michael ↔ Zue, Victor W.
(13)	Weischedel, Ralph M. ↔ Ayuso, Damaris M.
(13)	Manning, Christopher D. ↔ Klein, Dan
(13)	Rohlicek, J. Robin ↔ Ostendorf, Mari
(13)	Linebarger, Marcia C. ↔ Dahl, Deborah A.
(13)	Li, Wei ↔ Srihari, Rohini K.
(13)	Tanaka, Hozumi ↔ Tokunaga, Takenobu
(13)	Della Pietra, Stephen A. ↔ Della Pietra, Vincent J.
(13)	Seneff, Stephanie ↔ Polifroni, Joseph H.
(12)	Srihari, Rohini K. ↔ Niu, Cheng
(12)	Bobrow, Robert J. ↔ Ingria, Robert J. P.
(12)	Weischedel, Ralph M. ↔ Ramshaw, Lance A.
(12)	Niu, Cheng ↔ Li, Wei
(12)	Glass, James R. ↔ Phillips, Michael
(12)	Zue, Victor W. ↔ Polifroni, Joseph H.
(12)	Mercer, Robert L. ↔ Brown, Peter F.
(12)	Mercer, Robert L. ↔ Della Pietra, Vincent J.
(12)	Nagao, Makoto ↔ Tsujii, Jun'ichi
(12)	Zue, Victor W. ↔ Glass, James R.
(12)	Gale, William A. ↔ Church, Kenneth Ward
(12)	Grishman, Ralph ↔ Macleod, Catherine
(12)	Dahl, Deborah A. ↔ Norton, Lewis M.
(12)	Phillips, Michael ↔ Seneff, Stephanie

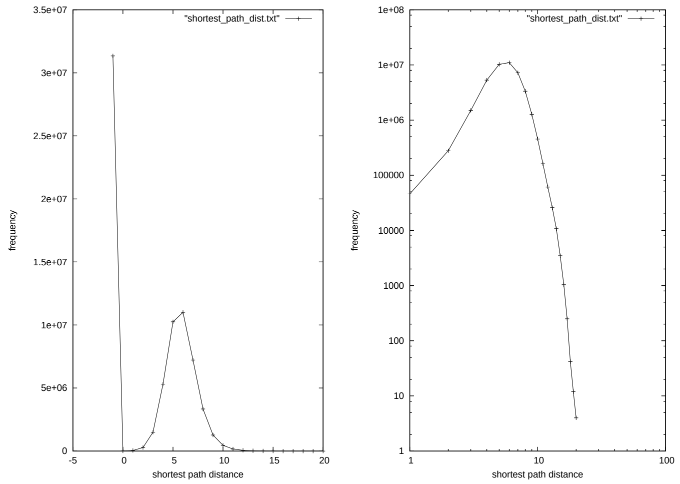
Table 20: *Author Collaboration Network Highest Edge Weights*

in both standard and loglog scale in Fig 14. It can clearly be seen that the shortest path distances follow a power law in the tail of the distribution. Newman's power law exponent is shown in Table 22.

The large number of disconnected author pairs is caused by the large number of connected components in the network. A lot of components in the graph with very few authors. But the sizes of the components are such that there is one giant connected component and all the other components are much smaller in size. The sizes of the components are listed in Table 23. The number of connected pairs of authors with a shortest path length of at most 6 contribute to 69% of the total number of connected pairs. Also, the power law in the tail of the distribution means that the components themselves are very tightly clustered, with the diameter of the network being just 20. This is a good sign in that the findings in the AAN research community will travel quickly through the community.

Shortest path distance	Frequency
-1	47783106
0	9421
1	45878
2	278636
3	1499786
4	5310534
5	10256634
6	110110580
7	226446
8	3341302
9	1275478
10	453042
11	161740
12	61094
13	25944
14	10794
15	3480
16	1038
17	250
18	42
19	12
20	4

Table 21: *Shortest path distance distribution in the Author Collaboration*



(a) Standard scale

(b) Logarithmic scale

Figure 14: *Shortest path distance distribution of the author collaboration network.*

α_{LS}	r^2	α_N	σ
3.11	0.86	2.61	0.0003

Table 22: *Power law measures for the shortest path distances in the author collaboration network*

Component Size	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	38	6400
Number of components	917	313	143	75	44	15	17	4	5	1	3	1	1	2	2	2	1	1	1	1

Table 23: *Sizes of the components in the author collaboration network*

6.2.4 Collaboration Network - PageRank

The PageRank centrality (Page et al., 1998) of the author collaboration network was computed. Both the weighted and unweighted network were analyzed and the results on the unweighted network can be seen in Table 24. Values are rounded to the nearest one hundred thousandth.

Both the weighted and the unweighted versions of the networks share generally the same central authors with 18 authors appearing in both lists.

6.2.5 Collaboration Network - Citation Network Centrality Correlation

In order to analyze the similaries between the author collaboration and citation networks, we calculated the correlation between the degree centrality values of authors in the collaboration network with those authors' scores in the citation network. Only authors that appeared in both networks were used for analysis.

We found that when all data points (9,421 authors) are included in the calculation, the Pearson correlation coefficient

Unweighted	
Author	**PageRank**
Tsujii, Jun'ichi	0.00091
Grishman, Ralph	0.00087
McKeown, Kathleen R.	0.00087
Hirschman, Lynette	0.00086
Palmer, Martha Stone	0.00085
Joshi, Aravind K.	0.00080
Wilks, Yorick	0.00079
Choi, Key-Sun	0.00078
Rambow, Owen	0.00077
Radev, Dragomir R.	0.00076
Weischedel, Ralph M.	0.00076
Matsumoto, Yuji	0.00074
Waibel, Alex	0.00073
Dagan, Ido	0.00072
Hovy, Eduard H.	0.00072
Huang, Chu-Ren	0.00071
Zhou, Ming	0.00069
Isahara, Hitoshi	0.00068
Marcu, Daniel	0.00066
Moore, Johanna D.	0.00065

Table 24: *Author Collaboration Network PageRanks*

is 0.68, a somewhat significant correlation. As the number of authors is reduced the correlation decreases dramatically. For instance, for the 200 authors with the highest collaboration degree centrality scores the correlation is reduced to 0.34. This seems to suggest that the most central authors in the collaboration network are not the same authors that are central to the citation network, yet there are a large number of authors who have low scores in both networks, which is to be expected. The results for the correlation coefficient for the top *n* authors can be seen in Table 25.

7 Conclusions

In this paper, we have analyzed statistically three networks composed from the citations between papers found in the ACL Anthology. These statistics include clustering coefficients, power law exponents, PageRank, and degree statistics.

All three networks display similar characteristics. Each one displays power law characteristics indicating a preference for edge attachment to a small number of high degree nodes, though the author collaboration shows a somewhat smaller. This shows that in each network, there are a small number of papers or authors which are attracting the majority of citations or collaborations.

# of Authors	Correlation Coefficient
50	0.03
100	0.28
200	0.34
500	0.49
1000	0.57
2000	0.62
5000	0.67
all(9421)	0.68

Table 25: *Correlation Coefficients Between Degree Centrality in Collaboration Network and Citation Citation Network*

Additionally, all of the networks display small world characteristics. This means that all of the networks are very well connected. This points to papers with many citations in the citation networks and a very active community of collaboration in the collaboration network.

We also observed that the author collaboration network is very tightly clustered and confirmed the existence of a power law in the tail of the shortest path lengths' distribution. This is good in the sense that new findings and ideas will propagate very quickly through the AAN research community.

All of the networks described show a strong tendency for certain authors and papers to play very strong roles in the overall structure of the network. Interestingly, the same authors do not occupy the same central positions in all of the networks, though there are several authors who consistently appear high in all ranked lists.

In addition to finding the most central papers and authors, we also analyzed the impact factor of journals, conferences and workshops. On analysis, it was observed that the impact factor of venues were increasing consistently over the past four decades. We also observed that conferences and workshops have shown a higher rate of growth in impact factor as compared to journals.

The maximum weighted edges in the author citation network are self citation edges. Although on further analysis, it is clear that the phenomenon of self-citation is not frequent enough in AAN to alter the rankings according to different measures of impact even slightly.

In our analysis, h-index does not appear to be strongly correlated with the number of incoming citations or PageRank. This is interesting, as the authors who have a high h-index also appear to have high incoming citations and PageRank. It is also clear that the h-index of an author in their subfield will differ, sometimes dramatically, from their overall h-index.

7.1 Future Work

We are currently pursuing the completion of a full textual statistical analysis of the papers composing the ACL Anthology Network. In particular we are looking into correlating lexical centrality and network centrality.

One factor we will investigate is LexRank. Recent research by (Erkan and Radev, 2004) applied centrality measures to assist text summarization. The system, LexRank, was successfully applied in the DUC 2004 evaluation, and was one of the top ranked systems in all four of the DUC 2004 Summarization tasks - achieving the best score in two of them. LexRank uses a cosine similarity adjacency matrix to identify predominant sentences of a text and then ranks these sentences according to centrality and salience. These groups of predominant sentences of individual papers could then be used to create another adjacency matrix between papers.

Additional factors we plan to investigate are the idea of ’Most Cited’ papers (Dervos and Kalkanis, 2005), self-citation (Fowler and Aksnes, 2007), and conference/venue specific impact factor.

Another area of interest is to build per-topic rankings (e.g., the most central papers in the area of statistical machine translation).

In the future, we also hope to expand our work by performing similar analysis for the PMCOA corpus and the SIGDA corpus.

The PMCOA, or PubMed Central Open Access, database is a free digital archive of journal articles in the biomedical and life sciences fields. It is maintained by the U.S. National Institutes of Health (NIH), and the papers in the Open Access list are mostly distributed under a Creative Commons license. More information can be found at their website (http://www.pubmedcentral.nih.gov/about/openftlist.html).

The SIGDA corpus is a collection of papers from the ACM Special Interest Group on Design Automation. It is a digital collection of papers dating back to 1989 from a number of different symposia, conferences, and journals - most notably, the ACM Transactions on Design Automation of Electronic Systems. More information can be found at their website (http://www.sigda.org/publications.html).

Also we are in the process of annotating the gender of all the authors. This annotated list will help us in further experiments on finding the correlation between gender and collaboration patterns.

Also we plan to rank venues based on the number of high quality publications that they have hosted. Instead of using all the publications with high incoming citations, we plan to use only publications which have been useful for increasing the h-index of an author.

Lastly, we plan to attempt to use network clustering techniques to categorize and label papers based on subject or topic, automatically. We anticipate that this categorization could help to highlight papers which might otherwise be missed in certain searches.

7.2 Acknowledgments

A number of students from the University of Michigan’s CLAIR Group helped with the work involved to create the data, network, and web pages. We would like to thank YoungJoo (Grace) Jeon, Mark Schaller, Ben Nash, John Umbaugh, Tunay Gur, Jahna Otterbacher, Arzucan Ozgur, Li Yang, Anthony Fader, Joshua Gerrish, Paul Hartzog, Chen Huang, Matt Simmons and Matt Masapollo.

Additional thanks go out to Brian Karrer and Professor Igor Markov for his assistance with ideas for this paper.

This work has been partially supported by the National Science Foundation grant ”Collaborative Research: BlogoCenter - Infrastructure for Collecting, Mining and Accessing Blogs”, jointly awarded to UCLA and the University of Michigan as IIS 0534323 to the University of Michigan and IIS 0534784 to UCLA and by the National Science Foundation grant ”iOPENER: A Flexible Framework to Support Rapid Learning in Unfamiliar Research Domains”, jointly awarded to the University of Maryland and the University of Michigan as IIS 0705832.

7.3 Availability of Data

The networks and associated metadata used in the analysis is available and can be downloaded from: http://belobog.si.umich.edu/clair/aan/downloads.cgi?package=aanrelease2007.tar.gz

References

Reka Albert and Albert-Laszlo Barabasi. 2001. Statistical mechanics of complex networks, Jun.

Mayur Amin and Michael Mabe. 2000. Impact factors: Use and abuse. *Perspectives in Publishing*, (1), October.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. Language Resources and Evaluation Conference, May.

Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel. 2006. Journal status. *Scientometrics*, 69:669.

Luciano da F. Costa, Francisco A. Rodrigues, Gonzalo Travieso, and P. R. Villas Boas. 2007. Characterization of complex networks: A survey of measurements.

Dimitris A. Dervos and Thomas Kalkanis. 2005. cc-IFF: A cascading citations impact factor framework for the automatic ranking of research publications. *IEEE International Workshop on Data Acquisition and Advanced Computing Systems Workshop Proceedings*, pages 668–673.

Sergey N. Dorogovtsev and Jose F. F. Mendes. 2002. Evolution of networks. *Advances in Physics*, 51:1079.

Nature Editorial. 2005. Not-so-deep impact. *Nature*, 435(7045):1003, June.

Ergin Elmacioglu and Dongwon Lee. 2005. On six degrees of separation in dblp-db and more. *SIGMOD Rec.*, 34(2):33–40.

Paul Erdös and Alfred Rényi. 1961. On the evolution of random graphs. *Bulletin of the International Statistics Institute*, 38:343–347.

Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, December 4.

Ramon Ferrer i Cancho and Ricard V. Solé. 2001. The small-world of human language. *Proceedings of the Royal Society of London B*, 268(1482):2261–2265, November 7.

James H. Fowler and Dag W. Aksnes. 2007. Does self-citation pay? *Scientometrics*, 72(3):427–437.

Anne-Wil Harzing. 2008. Publish or perish version 2.5.2969 (software). http://www.harzing.com/pop.htm visited June 18, 2008.

Jorge E. Hirsch. 2005. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102:16569.

Jorge E. Hirsch. 2007. Does the h-index have predictive power? *Proceedings of the National Academy of Sciences*, 104:19193.

Mark Thomas Joseph and Dragomir R. Radev. 2007. Citation analysis, centrality, and the ACL anthology. Technical Report CSE-TR-535-07, University of Michigan.

Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Steve Lawrence. 2001. Online or invisible? *Nature*, 411(6837):521.

Sune Lehmann, Andrew D. Jackson, and Benny E. Lautrup. 2006. Measures for measures. *Nature*, 444:1003–1004.

Elizabeth Leicht, Gavin Clarkson, Kerby Shedden, and Mark E. J. Newman. 2007. Large-scale structure of time evolving citation networks. *The European Physical Journal B*, 59:75.

Stanley Milgram. 1967. The small world problem. *Psychology Today*, pages 60–67, May.

Mario A. Nascimento, Jörg. Sander, and Jeff Pound. 2003. Analysis of SIGMODs coAuthorship graph. *ACM SIGMOD Record*, 32(3), September.

Mark E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review*, 45:167–265.

Mark E. J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, December.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Libary Technologies Project, January 29.

Dragomir R. Radev, Mark Hodges, Anthony Fader, Mark Joseph, Joshua Gerrish, Mark Schaller, Jonathan dePeri, and Bryan Gibson. 2007. Clairlib documentation v1.03. Technical Report CSE-TR-536-07, University of Michigan.

Erhard Rahm and Andreas Thor. 2005. Citation analysis of database publications. *ACM SIGMOD Record*, 34(4).

Xiaolin Shi, Belle Tseng, and Lada Adamic. 2008. Information diffusion in computer science citation networks. *Submitted*.

Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. 2006. Generalized h-index for disclosing latent facts in citation networks. http://arxiv.org/abs/cs.DL/0607066 visited on June 18, 2008.

B. Tjaden. 1996. The kevin bacon game.

Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, June 4.

Derek York. 1966. Least-square fitting of a straight line. *Canad. J. Phys.*, 44:1079–1086.