Citation Analysis, Centrality, and the ACL Anthology

Mark Thomas Joseph and Dragomir R. Radev
mtjoseph@umich.edu, radev@umich.edu
October 9, 2007
University of Michigan
Ann Arbor, MI 48109-1092

Abstract

We analyze the ACL Anthology citation network in an attempt to identify the most "central" papers and authors using graph-based methods. Citation data was obtained using text extraction from the library of PDF files with some post-processing performed to clean up the results. Manual annotation of the references was then performed to complete the citation network. The analysis compares metrics across publication years and venues, such as citations in and out. The most cited paper, central papers, and papers with the highest impact factor are also established.

1 Introduction

Bibliometrics is a popular method used to analyze paper and journal influence throughout the history of a work or publication. Statistically, this is accomplished by analyzing a number of factors, such as the number of times an article is cited.

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

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

For example, the impact factor over a two year period for a 2005 journal is equivalent to the 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).

Using network-based methods allowed us to also apply new methods to the analysis of a citation network, both textually and within the citation network. We applied a series of computations on the network, including LexRank and PageRank algorithms, as well as other measures of centrality and assorted network statistics.

Recent research by (Erkan and Radev, 2004) applied centrality measures to assist in the text summarization task. 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. We applied the LexRank system to the ACL citation network to identify central papers in the network based solely upon their textual content.

A significant amount of research has been devoted to published journal archives in past years. Recently a shift has been made to also statistically analyze the importance and significance of conference proceedings. Our research is an attempt to analyze not just journals and conferences, but to look at the entire history of an organization - the Association for Computational Linguistics (ACL). The ACL has been publishing a journal and sponsoring international conferences and workshops for over 40 years.

In the next section we review previous research into collaboration and citation networks, as well as summarize some of their findings. In section three, further information is provided regarding the contents of the ACL Anthology, an online repository of ACL's publishing history. The processing procedure is summarized in section four, including information on the text extraction, citation matching algorithm. The final sections cover both statistical and network computations of the ACL citation network.

2 Related Work

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. 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.

3 ACL Anthology

The Association for Computational Linguistics is an international and professional society dedicated to the advancement in Natural Language Processing and Computational Linguistics Research. The ACL Anthology is a collection of papers from an ACL published journal - Computational Linguistics - as well as all proceedings from ACL sponsored conferences and workshops.

Table 1 includes a listing of the different conferences and the meeting years we analyzed in Phase 1 of our 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 the 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. Also, more recent proceedings - most from 2007 - have been added. Finally, some of the missing proceedings of older years are now present. 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 the entire ACL Anthology is not included in this list - certain conference years are still being collected and archived, including the EACL-03 workshops and the proceedings of the 2007 conferences. Also, not every year has been completed, as articles from HLT-02 and COLING-65 are still absent.

Table 1: *ACL Conference Proceedings. This includes the years for which analysis was performed. Some years are still being collected and archived.*

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
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
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, 93-06
Computational Linguistics	J	74-05

In total, the ACL Anthology contains nearly 11,000 papers from these various sources, each with a unique ACL ID number. This number rises significantly if you include such listings as the Table of Contents, Front Matter, Author Indexes, Book Reviews, etc. For the sake of our work, these types of papers, and therefore these ACL IDs, have not been included in our computation.

Each of these papers was processed using OCR text extraction, and the references from each paper were parsed and extracted. These references were then manually matched to other papers in the ACL Anthology using an "n-best" (with $n = 5$) matching algorithm and a CGI interface. The manual annotation produced a citation network. The statistics of the anthology citation network in comparison to the total number of references in the 11,000 papers can be seen in Table 2.

Table 2: *General Statistics. A Citation is Considered Inside the Anthology if it Points to Another Paper in the ACL Anthology Network*

Total Papers Processed	10,921
Total Citations	152,546
Citations Inside Anthology	38,767, or approx. 25.4%
Citations Outside Anthology	113,779, or approx. 74.6%

4 Process

4.1 Metadata

A master list of ACL papers, authors, and venues was compiled using the data taken from the ACL Anthology website html. This metadata was stored in a simple text file in a format similar to BibTeX:

```
id = {}
author = {}
title = {}
year = {}
venue = {}
```

This file was used as the gold standard against which to match citations to their appropriate ACL ID numbers.

Post-processing was also performed on this metadata file. The accuracy of the information provided within the ACL webpages is impeccable, but in archiving 11,000 papers with the help of volunteers, mistakes are to be expected. Certain ACL IDs were mislabeled, with the corresponding PDF not matching the information provided. In other cases, author names were omitted or incorrectly identified.

One case that required a number of hours of manual cleanup was the consistency of author names. In attempting to build an author citation network and collaboration network to go along with the paper citation network, 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 webpages. Oftentimes, a comma was lost or missing to indicate the appropriate order of first and last name. Also, authors have a tendency to use different versions of their name over the course of their publishing career. For instance:

```
Michael Collins
Michael J. Collins
Michael John Collins
M. Collins
M. J. Collins
```

4.2 Text Extraction

The text extraction of the ACL Anthology was performed using PDFbox, an open source OCR text extraction program (`http://www.pdfbox.org/`). The contents of the ACL Anthology were extracted from the library of PDF's available from the repository hosted by the LDC. PDFbox was able to handle both one- and two-column papers layouts, making it ideal for the ACL Anthology which presents papers in both of these styles.

A separate script was written to find the "References/Bibliography/etc." section of each paper and to parse the individual references. After evaluating these results, it was determined that some pre-processing was necessary, as it was not uncommon for the "References" section to be split and for some references to be placed before the heading and/or within the body of a paper.

Other problems also surfaced. In one section of the ACL Anthology, namely the contents of the American Journal of Computational Linguistics Microfiche collections of 1974-1979, individual PDFs and ACL IDs actually represented collections of papers instead of a single paper. In this case, there could be several reference sections intermingled amongst approximately 100 pages of the PDF. In this case, the reference sections were manually extracted.

Also, the standards for PDF encoding have changed dramatically since its early inception, causing a number of the ACL papers - many of them older - to produce unusable or horribly jumbled text. To amend this problem, manual postprocessing was again performed. The references were either manually copied from these PDFs, or some cleaning was performed on the citation entries and return them to their original form.

Finally, because of the many different styles used in the past 40-plus years, the act of parsing references and identifying each individual references was difficult. To expedite the manual annotation process, the parsed reference results were manually examined and cleaned before the were passed to the annotation process.

4.3 Manual Annotation

The algorithm to match references from the ACL anthology to the gold standard was based on a simple keyword matching formula. Author, year, title, and venue were compared from the metadata against each reference. Comparisons scored a certain threshold of certainty, and the top five matches were returned.

These five matches were then presented to student researchers at the University of Michigan using a CGI interface. They were also provided with five additional options:

- Not Found - For those references that should have been found in the anthology but were not identified by the matching algorithm

- Related - For those references to non-ACL conference proceedings that share similar research interests (LREC, SIGIR, etc.)

- Not in Any - References not in the ACL Anthology or from related conference proceedings

- Unknown - For references extracted from PDFs with problematic encoding structures that were impossible to identify

- Not a Reference - For extra text that slipped past the manual annotator and did not represent an actual reference

It is estimated that for the 152,546 references in the 10,921 papers of the ACL Anthology, it took approximately 500 person-hours to complete the task. This evaluates to a little under 12 seconds for each reference.

4.4 The Networks

For our first network, we set each node to represent an ACL ID number, and the directed edges to represent a citation within that paper to the appropriate ID. For example then, the paper assigned ID no. P05-1002 results in the network in Table 3 and displayed in Figure 1. This network example includes the connections found between the papers cited by P05-1002. Additional stat

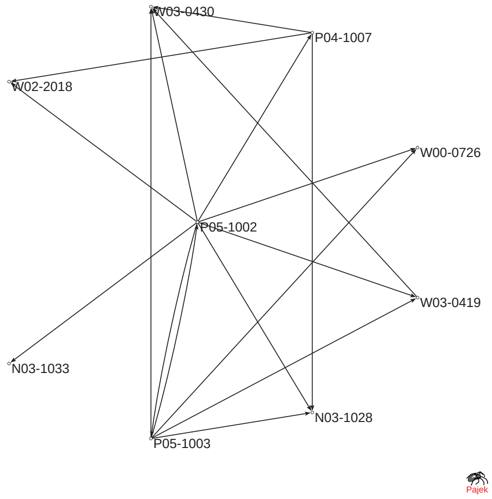

Figure 1: *Visual Representation of the Example Network Fragment for ACL ID no. P05-1002*

Next, basic statistics about the network, including most cited papers, outgoing citations per year, etc. were computed using a series of shell scripts. Impact analysis (as described above) was then computed manually using these statistics.

These same network calculations were also performed on the author citation network as well.

5 Statistical Results - Paper Network

Due to the size of the network, computation of certain factors in the network are time and resource intensive. In order to provide a picture of what the network looks like, we created and analyzed some smaller networks along with the full network. In this section you will find a breakdown of the statistics of these smaller networks and the full network.

As mentioned, the networks were analyzed using software from the University of Michigan CLAIR group. Some of the statistics you will see listed below are explained here.

The ACL Anthology Network is a directed network. A path between two nodes has a distance which is defined as the number of steps, or paths, that must be traversed to walk from one node to another. In larger or more dense graphs, numerous paths can be found from one node to another, and thus numerous distances exist between these two nodes. One common computation in network theory is known as the shortest path. The shortest path of a network is the shortest distance between two connected nodes. Two measures of shortest path were computed in our research. The first, developed by CLAIR, calculates the average of the shortest path between all vertices. The second comes from (Ferrer i Cancho and Solé, 2001), and is the average of all the average path lengths between the nodes.

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

“When the probability of measuring a particular value of some quantity varies inversely as a power of that value, the quantity is said to follow a power law, also known variously as Zipfs law or the Pareto distribution” (Newman, 2005). One of the ways to identify whether a network’s degree distribution demonstrates a power law relationship is to calculate the power law exponent (α) of the distribution. The accepted value of α that signifies a power law relationship is 2.5.

Here, power law exponents are calculated using two different methods. The first is through code developed by the CLAIR group, and is a measure of the slope of the cumulative log-log degree distribution. It is calculated as:

The power law exponent a is

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

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

r-squared r is

$$r = \frac{\sum xy}{\sqrt{(\sum xx * \sum yy)}}$$

where

$$\sum xy = \frac{(\sum(x*y)) - (\sum x * \sum y)}{n}$$

$$\sum xx = \frac{\sum x^2 - (\sum x)^2}{n}$$

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

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

Newman's power law exponent α is

$$\alpha = 1 + n[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}]^{-1}$$

where x_i and $i = 1...n$ are the measured values of x and x_{min} is again the minimum value of x

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

Newman's expected statistical error σ is

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

So, Newman's power law exponent for a network where

α = 2.500 and
σ = 0.002

would estimate to α = 2.500 $\pm$ 0.002.

The different power law measures were performed on the in-degree, out-degree, and total degree of the network. A table of the results for each of the networks can be found in their representative sections.

Finally, clustering coefficients are used to determine whether a network can be correctly identified as a small-world network. The ClairLib software calculates two types of clustering coefficient.

The first, Watts-Strogatz clustering coefficient, in (Watts and Strogatz, 1998), is computed as follows:

The clustering coefficient C is

$$C = \sum$$

- Watts-Strogatz clustering coefficient = 0.6243.
- Newman clustering coefficient = 0.4655.

The clustering coefficients here are significant, balancing nicely between a regular network and a random network. Thus it can be concluded that the network around P05-1002 is a Small World network.

5.2 TINLAP Only Network Characteristics

This network includes only the connection found between papers presented in the Proceedings of Theoretical Issues in Natural Language Processing (TINLAP). This was a small set of conferences that were held in 1975, 1978, and 1987. Any papers from outside venues and references/citations to or from those outside venues were removed. Power law exponent results can be found in Table 5.

- The TINLAP network consisted of 51 nodes, each representing a unique ACL ID number, and 50 directed edges.
- The diameter of the ACL Anthology Network graph is 4.
- The clairlib avg. directed shortest path: 1.62
- The Ferrer avg. directed shortest path: 0.99
- The harmonic mean geodesic distance: 41.76

Table 5: *TINLAP Network Power Law Measures*

Type of Degree	CLAIR Power Law	R-squared	Newman's Power Law	Newman's Error
in-degree	4.23	0.93	23.20	34.86
out-degree	2.21	0.98	2.77	0.74
total degree	2.58	0.99	3.75	1.02

Based on these values, the network does not appear to demonstrate a power law relationship under Newman's definition. The value of α is much higher than the expected 2.5 (here 3.75).

- Watts-Strogatz clustering coefficient = 0.0473.
- Newman clustering coefficient = 0.0426.

The clustering coefficients are both very low, thus it can be concluded that the TINLAP Network is not a Small World network.

5.3 ACL Only Network Characteristics

This network includes only the connection found between papers presented at the Annual Meeting of the Association for Computational Linguistics. Any papers from outside venues and references/citations to or from those outside venues were removed. Power law exponent results can be found in Table 6.

- The ACL-to-ACL network consisted of 1,541 nodes, each representing a unique ACL ID number, and 3,132 directed edges.
- The diameter of the ACL Anthology Network graph is 14.
- The clairlib avg. directed shortest path: 4.86

Table 6: *ACL-to-ACL Network Power Law Measures*

Type of Degree	CLAIR Power Law	R-squared	Newman's Power Law	Newman's Error
in-degree	2.76	0.94	2.57	0.08
out-degree	3.51	0.85	3.42	0.13
total degree	3.02	0.94	2.43	0.05

- The Ferrer avg. directed shortest path: 3.01
- The harmonic mean geodesic distance: 205.60

Based on these values, the network does appear to demonstrate a power law relationship under Newman's definition. The value of α is nearly 2.5 (here 2.43).

- Watts-Strogatz clustering coefficient = 0.1681.
- Newman clustering coefficient = 0.1361.

The clustering coefficients are both very low, thus it can be concluded that the entire ACL-to-ACL Network is not a Small World network.

5.4 Full Network Characteristics

This is the full ACL Anthology Network. It includes all connections found between ACL Anthology papers. Power law exponent results can be found in Table 7.

- The full network consisted of 8,898 nodes, each representing a unique ACL ID number, and 38,765 directed edges.
- The diameter of the ACL Anthology Network graph is 20.
- The clairlib avg. directed shortest path: 5.79
- The Ferrer avg. directed shortest path: 5.03
- The harmonic mean geodesic distance: 65.31

Table 7: *Full ACL Anthology Network Power Law Measures*

Type of Degree	CLAIR Power Law	R-squared	Newman's Power Law	Newman's Error
in-degree	2.54	0.97	2.03	0.02
out-degree	3.68	0.88	2.18	0.02
total degree	2.76	0.97	1.84	0.01

Based on these values, the network does not appear to demonstrate a full-blown power law relationship under Newman's definition. The value of α approaches 2.5, but is not statistically close enough.

- Watts-Strogatz clustering coefficient = 0.1878.
- Newman clustering coefficient = 0.0829.

The clustering coefficients of the full network are both very low, thus it can be concluded that the entire ACL Anthology Network is not a Small World network.

5.5 Anthology Statistics

Certain aspects of the anthology were analyzed quickly using shell scripts, yet these statistics still provide interesting insight into the ACL Anthology and the community. The 10 most cited papers within the anthology are listed in Table 8. Remember to refer to the prefix assignments for each conference and journal provided earlier to identify the year and venue of publication for each paper.

Table 8: *10 Most Cited Papers in the Anthology*

ACL ID	Title	Authors	Number of Times Cited
J93-2004	Building A Large Annotated Corpus Of English: The Penn Treebank	Mitchell P. Marcus; Mary Ann Marcinkiewicz; Beatrice Santorini	445
J93-2003	The Mathematics Of Statistical Machine Translation: Parameter Estimation	Peter F. Brown; Vincent J. Della Pietra; Stephen A. Della Pietra; Robert L. Mercer	344
J86-3001	Attention Intentions And The Structure Of Discourse	Barbara J. Grosz; Candace L. Sidner	308
A88-1019	Integrating Top-Down And Bottom-Up Strategies In A Text Processing System	Kenneth Ward Church	224
J96-1002	A Maximum Entropy Approach To Natural Language Processing	Adam L. Berger; Vincent J. Della Pietra; Stephen A. Della Pietra	188
A00-2018	A Classification Approach To Word Prediction	Eugene Charniak	184
P97-1003	Three Generative Lexicalized Models For Statistical Parsing	Michael John Collins	183
J95-4004	Transformation-Based-Error-Driven Learning And Natural Language Processing: A Case Study In Part-Of-Speech Tagging	Eric Brill	165
P95-1026	Unsupervised Word Sense Disambiguation Rivaling Supervised Methods	David Yarowsky	160
D96-0213	Figures Of Merit For Best-First Probabilistic Chart Parsing	Adwait Ratnaparkhi	160

The 10 papers with the largest numbers of references to other papers within the ACL Anthology Network are shown in Table 9. Because of this strong concentration on papers within the ACL Anthology Network, the assumption could be made that these papers are excellent examples of the types of research being done in the ACL community. This could be especially important for the present. With technology and research moving so quickly, it is refreshing to note that more than half of these papers have been published in the last 7 years. This is also a testament to the strength of the ACL Anthology as a research repository. Newer papers are referencing more and more papers within the anthology.

Further evidence that the number of citations in papers are rising can be seen in Table 10, where the most outgoing citations per year are calculated.

Table 11 shows the incoming citations by year, or the most cited years in the anthology - regardless of conference/journal. As expected, 2006 has yet to be cited, but 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. Conferences are seeing higher numbers of submissions and research continues to stay fresh and forward-thinking. Still, the unexplained 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 445 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 344 times) were both published in Computational Linguistics in 1993.

Table 9: *Papers with Most Citations within ACL Network*

ACL ID	Title	Authors	Number of References
J98-1001	Introduction To The Special Issue On Word Sense Disambiguation: The State Of The Art	Nancy M. Ide; Jean Veronis	59
J98-2002	Generalizing Case Frames Using A Thesaurus And The MDL Principle	Hang Li; Naoki Abe	38
J03-4003	Head-Driven Statistical Models For Natural Language Parsing	Michael John Collins	37
W06-2920	A Context Pattern Induction Method For Named Entity Extraction	Sabine Buchholz; Erwin Marsi	36
J00-4003	An Empirically Based System For Processing Definite Descriptions	Renata Vieira; Massimo Poesio	35
J05-1004	The Proposition Bank: An Annotated Corpus Of Semantic Roles	Martha Stone Palmer; Daniel Gildea; Paul Kingsbury	31
J93-2005	Lexical Semantic Techniques For Corpus Analysis	James D. Pustejovsky; Peter G. Anick; Sabine Bergler	31
J05-3002	Sentence Fusion For Multidocument News Summarization	Regina Barzilay; Kathleen R. McKeown	30
J05-3004	Comparing Knowledge Sources For Nominal Anaphora Resolution	Katja Markert; Malvina Nissim	30
W05-0620	Introduction To The CoNLL-2005 Shared Task: Semantic Role Labeling	Xavier Carreras; Lluis Marquez	30

Table 10: *Years with the Most Outgoing Citations*

Year	Outgoing Citations	Year	Outgoing Citations
2006	5765	1992	1327
2004	4430	1999	1316
2005	3812	1993	1069
2003	2732	1990	908
2000	2565	1991	796
2002	2506	1995	710
1998	2029	1988	592
1997	1791	1989	404
2001	1679	1986	339
1994	1529	1987	302
1996	1408	1984	183

Table 11: *Years with the Most Incoming Citations*

Year	Incoming Citations	Year	Incoming Citations
1993	2871	1990	1821
2002	2440	1995	1607
2000	2426	1999	1525
2003	2377	2001	1467
1998	2301	1988	1404
1997	2247	1991	1360
1992	2187	2005	1085
1996	2163	1986	1034
1994	2128	1989	930
2004	2028	1987	633

5.6 Impact Factor

Finally, impact factor was calculated for the ACL Anthology network based on a two year period using:

$$\frac{\text{Citations to Previous 2 Years}}{\text{No. of Articles Published in Previous 2 Years}}$$

The results can be found in Table 12 - rounded to the nearest thousandth.

Table 12: *Impact Factor for each Year*

Year	**Impact Factor**	**Year**	**Impact Factor**

Table 13: *Papers with the Highest PageRanks*

ACL ID	PageRank	Authors	Title
A88-1019	0.0229	Kenneth Ward Church	Integrating Top-Down And Bottom-Up Strategies In A Text Processing System
A88-1030	0.0188	Eva I. Ejerhed	The TIC: Parsing Interesting Text
C86-1033	0.0123	Geoffrey Sampson	A Stochastic Approach To Parsing
J90-2002	0.0097	Peter F. Brown; John Cocke; Stephen A. Della Pietra; Vincent J. Della Pietra; Frederick Jelinek; John D. Lafferty; Robert L. Mercer; Paul S. Roossin	A Statistical Approach To Machine Translation
P86-1022	0.0080	Joan Bachenko; Eileen Fitzpatrick; C. E. Wright	The Contribution Of Parsing To Prosodic Phrasing In An Experimental Text-To-Speech System
J86-3001	0.0073	Barbara J. Grosz; Candace L. Sidner	Attention Intentions And The Structure Of Discourse
J93-2004	0.0059	Mitchell P. Marcus; Mary Ann Marcinkiewicz; Beatrice Santorini	Building A Large Annotated Corpus Of English: The Penn Treebank
P83-1019	0.0049	Donald Hindle	Deterministic Parsing Of Syntactic Non-Fluencies
J93-2003	0.0045	Peter F. Brown; Vincent J. Della Pietra; Stephen A. Della Pietra; Robert L. Mercer	The Mathematics Of Statistical Machine Translation: Parameter Estimation
P84-1027	0.0045	Fernando C. N. Pereira; Stuart M. Shieber	The Semantics Of Grammar Formalisms Seen As Computer Languages
P83-1021	0.0042	Fernando C. N. Pereira; David H. D. Warren	Parsing As Deduction
C88-1016	0.0037	Peter F. Brown; John Cocke; Stephen A. Della Pietra; Vincent J. Della Pietra; Frederick Jelinek; Robert L. Mercer; Paul S. Roossin	A Statistical Approach To Language Translation
P84-1075	0.0035	Stuart M. Shieber	The Design Of A Computer Language For Linguistic Information
P83-1007	0.0034	Barbara J. Grosz; Aravind K. Joshi; Scott Weinstein	Providing A Unified Account Of Definite Noun Phrases In Discourse
P85-1018	0.0033	Stuart M. Shieber	Using Restriction To Extend Parsing Algorithms For Complex-Feature-Based Formalisms
P91-1034	0.0032	Peter F. Brown; Stephen A. Della Pietra; Vincent J. Della Pietra; Robert L. Mercer	Word-Sense Disambiguation Using Statistical Methods
J92-4003	0.0031	Peter F. Brown; Peter V. DeSouza; Robert L. Mercer; Thomas J. Watson; Vincent J. Della Pietra; Jennifer C. Lai	Class-Based N-Gram Models Of Natural Language
J88-1003	0.0030	Steven J. DeRose	Grammatical Category Disambiguation By Statistical Optimization
J81-4003	0.0030	Fernando C. N. Pereira	Extraposition Grammars
P82-1028	0.0029	Kathleen R. McKeown	The Text System For Natural Language Generation: An Overview

Table 14: *Papers with the Highest PageRanks per Year*

ACL ID	PageRank per Year	Authors	Title
A88-1019	0.00115	Kenneth Ward Church	Integrating Top-Down And Bottom-Up Strategies In A Text Processing System
A88-1030	0.00099	Eva I. Ejerhed	The TIC: Parsing Interesting Text
C86-1033	0.00057	Geoffrey Sampson	A Stochastic Approach To Parsing
J90-2002	0.00057	Peter F. Brown; John Cocke; Stephen A. Della Pietra; Vincent J. Della Pietra; Frederick Jelinek; John D. Lafferty; Robert L. Mercer; Paul S. Roossin	A Statistical Approach To Machine Translation
J93-2004	0.00042	Mitchell P. Marcus; Mary Ann Marcinkiewicz; Beatrice Santorini	Building A Large Annotated Corpus Of English: The Penn Treebank
P86-1022	0.00038	Joan Bachenko; Eileen Fitzpatrick; C. E. Wright	The Contribution Of Parsing To Prosodic Phrasing In An Experimental Text-To-Speech System
J86-3001	0.00035	Barbara J. Grosz; Candace L. Sidner	Attention Intentions And The Structure Of Discourse
J93-2003	0.00032	Peter F. Brown; Vincent J. Della Pietra; Stephen A. Della Pietra; Robert L. Mercer	The Mathematics Of Statistical Machine Translation: Parameter Estimation
J96-1002	0.00023	Adam L. Berger; Vincent J. Della Pietra; Stephen A. Della Pietra	A Maximum Entropy Approach To Natural Language Processing
J02-3001	0.00021	Daniel Gildea; Daniel Jurafsky	Automatic Labeling Of Semantic Roles
J92-4003	0.00021	Peter F. Brown; Peter V. DeSouza; Robert L. Mercer; Thomas J. Watson; Vincent J. Della Pietra; Jennifer C. Lai	Class-Based N-Gram Models Of Natural Language
P83-1019	0.00020	Donald Hindle	Deterministic Parsing Of Syntactic Non-Fluencies
P91-1034	0.00020	Peter F. Brown; Stephen A. Della Pietra; Vincent J. Della Pietra; Robert L. Mercer	Word-Sense Disambiguation Using Statistical Methods
P84-1027	0.00020	Fernando C. N. Pereira; Stuart M. Shieber	The Semantics Of Grammar Formalisms Seen As Computer Languages
C88-1016	0.00020	Peter F. Brown; John Cocke; Stephen A. Della Pietra; Vincent J. Della Pietra; Frederick Jelinek; Robert L. Mercer; Paul S. Roossin	A Statistical Approach To Language Translation
P02-1040	0.00019	Kishore Papineni; Salim Roukos; Todd Ward; Wei-Jing Zhu	Bleu: A Method For Automatic Evaluation Of Machine Translation
P91-1022	0.00018	Peter F. Brown; Jennifer C. Lai; Robert L. Mercer	Aligning Sentences In Parallel Corpora
D96-0213	0.00018	Adwait Ratnaparkhi	Figures Of Merit For Best-First Probabilistic Chart Parsing
A00-2018	0.00018	Eugene Charniak	A Classification Approach To Word Prediction
P83-1021	0.00018	Fernando C. N. Pereira; David H. D. Warren	Parsing As Deduction

Table 15: *Repeated Top PageRank Papers*

ACL ID	In-Degree	Out-Degree	Total Edges	Percent
A88-1019	224	1	225	0.58
A88-1030	5	2	7	0.02
C86-1033	9	0	9	0.02
J90-2002	142	1	143	0.37
P86-1022	4	0	4	0.01
J86-3001	308	6	314	0.81
J93-2004	445	8	453	1.17
J93-2003	344	8	352	0.91
P83-1019	36	3	39	0.10
P84-1027	20	5	25	0.06
P83-1021	44	3	47	0.12
C88-1016	26	1	27	0.07
P91-1034	66	2	68	0.18
J92-4003	130	1	131	0.34
Total	1,803	41	1,844	4.76
Full Network	38,765 total edges			

changes in rank. In Table 18, we list the changes of the ACL IDs found in the top 20 PageRank and PageRank per Year charts.

7 Results - Author Networks

Because much research has been published regarding the networks formed by author interactions in a digital collection we created both an author citation network and an author collaboration network. The following two sections describe in greater detail these two networks, as well as provide statistics and comparisons to other research. A number of statistical measures were performed, including centrality, clustering coefficients, PageRank, and degree statistics.

7.1 Citation Network

The ACL Anthology author citation network is based on the ACL Anthology Network. Here though, one author cites another author. So for any paper, each author of that paper would occur as a node in the network. If this ACL Anthology paper were to cite another ACL Anthology paper, then the author(s) of the first paper would cite the author(s) of the second paper. For a more concrete example: if Hal Daume III writes an ACL Anthology paper and cites an earlier work by James D. Pustejovsky, then the link "Daume III, Hal → Pustejovsky, James D." would occur in the network. Also, we have decided to include self-citation in the network.

As stated earlier, a number of measures were calculated for this network. We start with some general statistics, centrality and clustering coefficients. Power law exponent results can be found in Table 19.

7.2 Citation Network - Centrality and Clustering Coefficients

- The Author Citation Network consisted of 7,090 nodes, each representing a unique author, and 137,007 directed edges.
- The diameter of the Author Citation Network graph is 9.
- The clairlib avg. directed shortest path: 3.35
- The Ferrer avg. directed shortest path: 3.32
- The harmonic mean geodesic distance: 5.42

Table 16: *Top Gainers in PageRank Normalization*

ACL ID	PageRank Rating	PageRank/Year Rating	Gain
N06-1057	8895	1407	+7488
P06-1125	8893	1406	+7487
P06-1105	8868	1403	+7465
P06-1118	8869	1404	+7465
E06-1023	8870	1405	+7465
P06-2043	8866	1402	+7464
W06-1708	8863	1401	+7462
W06-1413	8847	1400	+7447
P06-1147	8841	1399	+7442
W06-1516	8839	1398	+7441
P06-1073	8832	1397	+7435
P06-4001	8830	1396	+7434
P06-2090	8828	1395	+7433
W06-1703	8825	1393	+7432
N06-1005	8826	1394	+7432
P06-2021	8820	1392	+7428
W06-1002	8816	1390	+7426
W06-0507	8817	1391	+7426
P06-2051	8806	1389	+7417
W06-2809	8802	1388	+7414
W06-0907	8799	1387	+7412
P06-2005	8792	1386	+7406
W06-2205	8784	1384	+7400
W06-2907	8785	1385	+7400
W06-1203	8770	1382	+7388
E06-1051	8771	1383	+7388
P06-3015	8760	1379	+7381
N06-2020	8761	1380	+7381
W06-0122	8762	1381	+7381
D06-1611	8758	1378	+7380

Table 17: *Top Losers in PageRank Normalization*

ACL ID	PageRank Rating	PageRank/Year Rating	Loss
J79-1047	1872	7405	-5533
J79-1036f	1871	7404	-5533
P79-1016	2575	8121	-5546
J79-1044	2146	7694	-5548
C73-2025	1158	6732	-5574
T75-2027	2917	8509	-5592
T78-1026	1866	7459	-5593
T78-1027	1862	7457	-5595
C69-6801	3117	8722	-5605
C69-2001	3084	8721	-5637
C69-1801	3054	8720	-5666
C69-1401	3041	8719	-5678
C69-0201	3039	8718	-5679
T78-1006	2117	7802	-5685
C65-1021	3105	8791	-5686
C67-1023	3079	8766	-5687
T78-1014	2112	7799	-5687
C67-1025	3055	8765	-5710
C65-1014	3037	8790	-5753
C73-2019	2830	8585	-5755
C67-1020	951	6736	-5785
C67-1002	950	6735	-5785
T75-2008	1772	7616	-5844
T75-2014	1928	7821	-5893
C67-1007	2628	8640	-6012
C65-1024	2152	8498	-6346

Table 18: *Movement of Top PageRanks Due to Normalization*

ACL ID	PageRank Rating	PageRank/Year Rating	Change
A88-1019	1	1	0
A88-1030	2	2	0
C86-1033	3	3	0
J90-2002	4	4	0
P86-1022	5	6	-1
J86-3001	6	7	-1
J93-2004	7	5	+2
P83-1019	8	12	-4
J93-2003	9	8	+1
P84-1027	10	14	-4
P83-1021	11	20	-9
C88-1016	12	15	-3
P84-1075	13	27	-14
P83-1007	14	32	-18
P85-1018	15	29	-14
P91-1034	16	13	+3
J92-4003	17	11	+6
J88-1003	18	23	-5
J81-4003	19	45	-26
P82-1028	20	42	-22
J96-1002	25	9	+16
J02-3001	108	10	+98
P02-1040	127	16	+111
P91-1022	21	17	+4
D96-0213	42	18	+24
A00-2018	88	19	+69

Table 19: *Author Citation Network Power Law Measures*

Type of Degree	CLAIR Power Law	R-squared	Newman's Power Law	Newman's Error
in-degree	2.22	0.91	1.57	0.01
out-degree	2.59	0.84	1.56	0.01
total degree	2.29	0.89	1.47	0.00

Based on these values, the network not does appear to demonstrate a power law relationship under Newman's definition. The value of α is too low in comparison to the expected 2.5 (here 1.47).

- Watts-Strogatz clustering coefficient = 0.4702.
- Newman clustering coefficient = 0.1484.

The Wattz-Strogatz clustering coefficient is nearly 0.5, therefore the author citation network could be considered a Small World Network. On the other hand, the Newman clustering coefficient is much too low, thus it can be concluded that the network is not a Small World network according to Newman.

7.3 Citation Network - Degree Statistics

In Table 20, we show the top 20 authors for both in-coming and out-going citations. Out-going citations refer to the number of times an author cites other authors within the ACL Anthology. In-coming citations refer to the most cited authors within the ACL Anthology.

Table 20: *Author Citation Network Highest In- and Out-Degrees*

Out-Degree		In-Degree	
(1144)	Ney, Hermann	(2302)	Della Pietra, Vincent J.
(977)	Tsujii, Jun'ichi	(2136)	Mercer, Robert L.
(950)	McKeown, Kathleen R.	(2097)	Church, Kenneth Ward
(886)	Marcu, Daniel	(2029)	Della Pietra, Stephen A.
(789)	Grishman, Ralph	(1933)	Marcus, Mitchell P.
(757)	Matsumoto, Yuji	(1920)	Brown, Peter F.
(676)	Joshi, Aravind K.	(1897)	Och, Franz Josef
(675)	Hovy, Eduard H.	(1798)	Ney, Hermann
(645)	Palmer, Martha Stone	(1608)	Collins, Michael John
(639)	Collins, Michael John	(1516)	Yarowsky, David
(628)	Lapata, Maria	(1328)	Brill, Eric
(568)	Carroll, John A.	(1289)	Joshi, Aravind K.
(563)	Weischedel, Ralph M.	(1270)	Santorini, Beatrice
(555)	Hirschman, Lynette	(1266)	Marcinkiewicz, Mary Ann
(550)	Poesio, Massimo	(1259)	Charniak, Eugene
(549)	Gildea, Daniel	(1211)	Pereira, Fernando C. N.
(544)	Wiebe, Janyce M.	(1208)	Grishman, Ralph
(532)	Knight, Kevin	(1099)	Grosz, Barbara J.
(531)	Manning, Christopher D.	(1067)	Knight, Kevin
(528)	Johnson, Mark	(1062)	Roukos, Salim

In Table 21, the top 30 weighted edges are listed from the citation network. The weight is the edge weight, which represents the number of times one author citing another occurs. So, for instance, as you can see from the chart, Hermann Ney cites different works by Franz Josef Och 103 times. Remember that individual papers could have multiple references to papers by the same author.

Although not surprising, as it is common to cite your own research, it is still noteworthy that 21 of the top 30 strongest edges in the graph are self-citations. This shows not only the importance 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 with more papers written over a period of time because of those author's self-citations.

7.4 Citation Network - PageRank

Finally, the PageRank centrality of the author citation network was computed. For this situation, 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 is as described above, whereas the unweighted network treats all multiple incidents of a citation as a single occurrence.

Table 21: *Author Citation Network Highest Edge Weights*

(145)	Ney, Hermann → Ney, Hermann
(103)	Ney, Hermann → Och, Franz Josef
(78)	Joshi, Aravind K. → Joshi, Aravind K.
(77)	Grishman, Ralph → Grishman, Ralph
(74)	Tsujii, Jun'ichi → Tsujii, Jun'ichi
(67)	Ney, Hermann → Della Pietra, Vincent J.
(66)	Ney, Hermann → Della Pietra, Stephen A.
(66)	Ney, Hermann → Tillmann, Christoph
(65)	Seneff, Stephanie → Seneff, Stephanie
(61)	Och, Franz Josef → Ney, Hermann
(60)	Weischedel, Ralph M. → Weischedel, Ralph M.
(58)	Ney, Hermann → Mercer, Robert L.
(58)	Ney, Hermann → Brown, Peter F.
(57)	Litman, Diane J. → Litman, Diane J.
(56)	McKeown, Kathleen R. → McKeown, Kathleen R.
(52)	Johnson, Mark → Johnson, Mark
(51)	Schabes, Yves → Schabes, Yves
(51)	Palmer, Martha Stone → Palmer, Martha Stone
(49)	Och, Franz Josef → Och, Franz Josef
(49)	Knight, Kevin → Knight, Kevin
(47)	Bangalore, Srinivas → Bangalore, Srinivas
(47)	Zue, Victor W. → Seneff, Stephanie
(46)	Poesio, Massimo → Poesio, Massimo
(46)	Wu, Dekai → Wu, Dekai
(46)	Rambow, Owen → Rambow, Owen
(46)	Hovy, Eduard H. → Hovy, Eduard H.
(45)	Zens, Richard → Ney, Hermann
(45)	Harabagiu, Sanda M. → Harabagiu, Sanda M.
(44)	Wiebe, Janyce M. → Wiebe, Janyce M.
(44)	Schwartz, Richard M. → Schwartz, Richard M.

The top weighted and unweighted PageRank results can be seen in Table 22. Please note the values have been rounded.

Table 22: *Author Citation Network PageRanks*

Weighted		Unweighted	
Author	**PageRank**	**Author**	**PageRank**
Church, Kenneth Ward	0.00936	Mercer, Robert L.	0.01413
Della Pietra, Vincent J.	0.00651	Church, Kenneth Ward	0.01391
Sampson, Geoffrey	0.00613	Della Pietra, Vincent J.	0.01257
Della Pietra, Stephen A.	0.00605	Brown, Peter F.	0.01211
Mercer, Robert L.	0.00601	Della Pietra, Stephen A.	0.01164
Brill, Eric	0.00576	Sampson, Geoffrey	0.00954
Marcus, Mitchell P.	0.00570	Jelinek, Frederick	0.00851
Brown, Peter F.	0.00541	Marcus, Mitchell P.	0.00849
Pereira, Fernando C. N.	0.00521	Brill, Eric	0.00671
Grosz, Barbara J.	0.00505	Weischedel, Ralph M.	0.00629
Jelinek, Frederick	0.00480	Joshi, Aravind K.	0.00581
Hindle, Donald	0.00474	Lafferty, John D.	0.00580
Joshi, Aravind K.	0.00450	Grosz, Barbara J.	0.00578
Weischedel, Ralph M.	0.00440	Pereira, Fernando C. N.	0.00572
Gale, William A.	0.00432	Hindle, Donald	0.00557
Santorini, Beatrice	0.00408	Santorini, Beatrice	0.00549
Lafferty, John D.	0.00390	Gale, William A.	0.00504
Sidner, Candace L.	0.00374	Roossin, Paul S.	0.00502
Grishman, Ralph	0.00374	Cocke, John	0.00502
Roukos, Salim	0.00356	Schwartz, Richard M.	0.00490

Both weighted and unweighted networks still generally share the same central authors in the ACL Citation Network - with only 3 out of 20 unique authors in comparison.

7.5 Collaboration Network

The ACL Anthology author collaboration network is based on the metadata of the ACL Anthology. Whenever one author co-authors (or collaborates) with another author, a vector between the two is formed. For instance, ACL ID N04-1005 refers to “Balancing Data-Driven And Rule-Based Approaches In The Context Of A Multimodal Conversational System” by Srinivas Bangalore and Michael Johnston. This would create the vector “Bangalore, Srinivas $\leftrightarrow$ Johnston, Michael” in the network. Because of the nature of a collaboration, it should be noted that this network is undirected.

As stated earlier, a number of measures were calculated for this network. We start with some general statistics, centrality and clustering coefficients. Power law exponent results can be found in Table 23. Note that because this network is undirected, only the total degree power law measure has been included.

7.6 Collaboration Network - Centrality and Clustering Coefficients

- The Author Collaboration Network consisted of 7,854 nodes, each representing a unique author, and 41,370 directed edges.
- The diameter of the Author Collaboration Network graph is 17.
- The clairlib avg. directed shortest path: 6.04
- The Ferrer avg. directed shortest path: 4.69
- The harmonic mean geodesic distance: 10.15

Note the average directed shortest path as calculated in with ClairLib software is 6.04. This nearly mirrors (Milgram, 1967)’s “six degrees of separation” experiments.

Table 23: *Author Collaboration Network Power Law Measure*

ClairLib Power Law	3.15
R-squared	0.90
Newman's Power Law	1.81
Newman's Error	0.01

Based on the value, the network may demonstrate a power law relationship under Newman's definition, but not a strong one. The value of α is lower than the expected 2.5 (here 1.81).

- Watts-Strogatz clustering coefficient = 0.6341.

- Newman clustering coefficient = 0.3952.

The Wattz-Strogatz clustering coefficient is above 0.5, therefore the author collaboration network should be considered a Small World Network. The Newman clustering coefficient approaches 0.5, thus it can be concluded that the network is almost a Small World network according to Newman.

How does this compare to other research and other digital collections? The results of other research is included in comparison to our findings for the ACL Anthology Network in Table 24. Please note that the results from other research may not include matching algorithms used to find certain values. Labels have been made as specific as possible. When the method used to find a value in other research is not found, the value is placed across both categories.

Table 24: *Author Collaboration Networks - Statistics*

	Power Law Exponent		**Clustering Coefficient**	
Archive	**ClairLib**	**Newman's**	**Watts-Strogatz**	**Newman**
DBLP (Elmacioglu and Lee, 2005)	3.68		0.63	
ACL Anthology (this paper)	3.15	0.90	0.6341	0.3952

7.7 Collaboration Network - Degree Statistics

In Table 25, we show the top 20 authors with the most collaborations in the ACL Anthology Network, with the number of collaboration they have been party to.

Table 25: *Author Collaboration Network Most Collaborations*

(171)	Tsujii, Jun'ichi	(102)	McKeown, Kathleen R.
(167)	Hirschman, Lynette	(101)	Waibel, Alex
(165)	Weischedel, Ralph M.	(100)	Ney, Hermann
(156)	Schwartz, Richard M.	(100)	Palmer, Martha Stone
(151)	Isahara, Hitoshi	(98)	Roukos, Salim
(123)	Joshi, Aravind K.	(96)	Seneff, Stephanie
(118)	Grishman, Ralph	(96)	Matsumoto, Yuji
(113)	Wilks, Yorick	(92)	Zue, Victor W.
(112)	Ingria, Robert J. P.	(91)	Makhoul, John
(110)	Rayner, Manny	(90)	Lavie, Alon

In Table 26, the top 34 weighted edges are listed from the collaboration network. The weight is the edge weight, which represents the number of times the two authors have collaborated together. So, for instance, as you can see from the chart, Yusuke Miyao has co-authored 20 papers with Jun'ichi Tsujii.

Table 26: *Author Collaboration Network Highest Edge Weights*

(21)	Makhoul, John	↔	Schwartz, Richard M.
(20)	Tsujii, Jun'ichi	↔	Miyao, Yusuke
(18)	Uchimoto, Kiyotaka	↔	Isahara, Hitoshi
(17)	Murata, Masaki	↔	Isahara, Hitoshi
(17)	Joshi, Aravind K.	↔	Webber, Bonnie Lynn
(16)	Isahara, Hitoshi	↔	Ma, Qing
(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)	Lin, Chinyew	↔	Hovy, Eduard H.
(14)	Grishman, Ralph	↔	Sterling, John
(13)	Rayner, Manny	↔	Hockey, Beth Ann
(13)	Phillips, Michael	↔	Zue, Victor W.
(13)	Weischedel, Ralph M.	↔	Ayuso, Damaris M.
(13)	Manning, Christopher D.	↔	Klein, Dan
(13)	Zens, Richard	↔	Ney, Hermann
(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)	Wu, Dekai	↔	Carpuat, Marine
(12)	Glass, James R.	↔	Phillips, Michael
(12)	Zue, Victor W.	↔	Polifroni, Joseph H.
(12)	Mercer, Robert L.	↔	Brown, Peter F.
(12)	Della Pietra, Vincent J.	↔	Mercer, Robert L.
(12)	Nagao, Makoto	↔	Tsujii, Jun'ichi

7.8 Collaboration Network - PageRank

Lastly, the PageRank centrality of the author collaboration network was computed. For this situation, in order to avoid bias due to repeated collaborations, we analyzed two different networks, both an unweighted and a weighted collboration network. The weighted network is as described above, whereas the unweighted network treats all multiple incidents as a single occurrence.

The top weighted and unweighted PageRank results can be seen in Table 27. Please note the values have been rounded.

Table 27: *Author Collaboration Network PageRanks*

Weighted		Unweighted	
Author	**PageRank**	**Author**	**PageRank**
Tsujii, Jun'ichi	0.00099	Tsujii, Jun'ichi	0.00147
Hirschman, Lynette	0.00094	Joshi, Aravind K.	0.00125
Wilks, Yorick	0.00086	Isahara, Hitoshi	0.00112
McKeown, Kathleen R.	0.00085	Hirschman, Lynette	0.00110
Joshi, Aravind K.	0.00085	Weischedel, Ralph M.	0.00106
Choi, Key-Sun	0.00084	McKeown, Kathleen R.	0.00105
Weischedel, Ralph M.	0.00084	Wilks, Yorick	0.00104
Waibel, Alex	0.00083	Matsumoto, Yuji	0.00097
Matsumoto, Yuji	0.00079	Grishman, Ralph	0.00096
Radev, Dragomir R.	0.00077	Waibel, Alex	0.00095
Huang, Chu-Ren	0.00075	Choi, Key-Sun	0.00095
Isahara, Hitoshi	0.00075	Palmer, Martha Stone	0.00089
Grishman, Ralph	0.00075	Moldovan, Dan I.	0.00089
Palmer, Martha Stone	0.00075	Huang, Chu-Ren	0.00084
Rambow, Owen	0.00071	Rambow, Owen	0.00084
Marcu, Daniel	0.00071	Nagao, Makoto	0.00084
Strzalkowski, Tomek	0.00070	Radev, Dragomir R.	0.00082
Shriberg, Elizabeth	0.00070	Ney, Hermann	0.00082
Dorr, Bonnie Jean	0.00067	Huang, Changning	0.00081
Dagan, Ido	0.00066	Nirenburg, Sergei	0.00079

Both weighted and unweighted networks generally share the same central authors in the ACL Collaboration Network - with only 5 out of 20 unique authors in comparison.

8 Conclusions

In this paper, we have statistically analyzed a number of different factors in the ACL Anthology Network. This includes clustering coefficients, power law exponents, PageRank, and degree statistics.

In comparison to other research performed in bibliometrics applied to large digital collections, the ACL Anthology Network displays some interesting behavior. We have summarized some of the important statistics from our analysis and combined them with other research.

9 Future Work

We are currently pursuing the completion of a full statistical analysis of the ACL Anthology Network. Because of the size of the network, the processing time required to analyze not just a network of this size but also the full text of those articles is large. We are also looking into methods for calculating h-index and a conference/venue specific impact factor.

Clustering methods are also going to be performed in the hopes of classifying texts by subject. We hope this form of community finding will lead to renewed interests in certain papers, and work as a knowledge source for authors and researchers in different aspects of Natural Language Processing.

Also, we hope to release the fruits of our labor to the public for future research purposes.

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 biomedial 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`).

Lastly, we plan to implement some form of network clustering in the hopes discovering new ways to categorize and label papers based on subject or topic using only graph based algorithms.

10 Acknowledgments

A number of students from the University of Michigan's CLAIR Group helped with the work involved to create the data, network, and webpages. 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, and Bryan Gibson.

A special thanks goes out to University of Michigan 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 UMich as IIS 0534323 to UMich 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 UMd and UMich as IIS 0705832.

References

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

Ergin Elmacioglu and Dongwon Lee. 2005. On six degrees of separation in DBLP-DB and more. *ACM SIGMOD Record*, 34(2):33–40.

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.

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. *Sigmod Record*, 32(3), September.

Mark E. J. Newman, Duncan J. Watts, and S. H. Strogatz. 2002. Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:2566–2572, February. Suppl.1.

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

L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Libary Technologies Project, January 29.

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

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

Appendix: Release notes

The following is a copy of a report made to members of the LDC (`http://www.ldc.upenn.edu/`) and the dAnth group (`http://wing.comp.nus.edu.sg/mailman/listinfo/dAnth/`), two groups involved and interested in the ACL Anthology collection. It is printed here nearly verbatim, with some omissions of names and format changes to improve layout. It can be used for further explanation regarding some of the inconsistencies involved in such a large collection of electronic documents.

In response to some of the questions posed to the authors, and in an attempt to document some of the foibles I encountered while working with ACL anthology, we have compiled this list of different problems with the ACL Anthology as it is currently presented online. We are working here with the most recent version, as hosted at `http://acl.ldc.upenn.edu/`. We apologize if any of this information is redundant.

Please feel free to direct any further questions you may have to the authors via email. We will do our best to expound on the contents of this report or regarding any of these questions.

I have divided this report into the following sections:

1. The TGZ Files == regarding the downloadable archives of the contents of the ACL Anthology
2. The Website == regarding the information contained on the website
3. The Papers == regarding the actual PDF versions of the papers
4. Other == other thoughts and issues that do not fall cleanly under the previous three

10.1 The TGZ Files

The following IDs are included in the tgz files, but are duplicates due to two conferences being held in conjunction. The IDs in parentheses are the equivalent papers included in the anthology and already included in the tgz files as well. We do not know if this is an intentional method intended to allow visitors to download only one conference's proceedings. But, if that is the case, then there should be more incidences of this overlap because of the number of conferences that have been held jointly.

- C98-1000 to C98-1117 (P98-1000 to P98-1117)
- C98-2000 (P98-2000)
- C98-2118 to C98-2246 (P98-2123 to P98-2151)
- E97-1000 to E97-1073 (P97-1000 to P97-1073)

The following ids are missing from the tgz files, but they are listed on the website.

- E03-1062
- E03-1063
- E03-1082
- E03-1083
- I05-all
- W01-0704

- W01-0705
- W01-0708
- W01-0711
- W01-0720
- W01-0721
- W01-0722
- W01-0724
- W01-0725
- W01-1018
- W01-1310

The following IDs and their pdf counterparts do not have matching names. The actual name is followed by the pdf file name in parentheses. This is also a problem because the webpages are encoded to link to the correct name, which leads to a person being provided with a multiple choice of options for matching documents.

- N04-2001 (N04-2-01)
- N04-2002 (N04-2-02)
- N04-2003 (N04-2-03)
- N04-2004 (N04-2-04)
- N04-2005 (N04-2-05)
- N04-2006 (N04-2-06)
- N04-2007 (N04-2-07)
- N04-2008 (N04-2-08)
- N04-2009 (N04-2-09)
- N04-2010 (N04-2-10)

Also, the W04- set comes also with a series of files entitled ".Zap.*" where the star represents some ACL ID from the W04- collection. So, for instance, there is a ".Zap.W04-1001.pdf" file. We are not sure if these have a specific purpose.

10.2 The Website

Both C86-1062 and C86-1065 are labeled as the same paper on the website, but C86-1065 should be ”A Morphological Recognizer with Syntactic and Phonological Rules” by John Bear.

The listings for the H05- set are not in ACL ID number order. H05-1011 thru H05-1099 are located at the end of the page.

There are also a large number of misspellings, omissions, and misordered (last name first) author names on the webpages. Here is a short listing of some of the author problems. It might be worth considering standardizing the author names if this to be released as a corpus. The name as it appears is first, and then in parentheses is the assumed fix if available.

- Yuji Matsumo (Yuji Matsumoto)
- Yuka Tateishi (Yuka Tateisi)
- Yung-Taek Kim/Yung Taek Kim/Yung Tack Kim (three different uses)
- Zoyn M. Shalyapina (Zoya, not Zoyn)
- Youn S. Han (Young S. Han)
- Yoshimi Suzukit (Suzuki, not Suzukit - this often happens when the name is labelled with a footnote in the shape of a cross)
- Yoshilco Lto
- Anne Demerits (Demedts)
- Tailco Dietzel (Taiko)
- E. Jelinek (F.)
- Klein Dan (switch)
- Yang (2) Liu and Yang (1) Liu (For some reason, the (1) and (2) appear in line)
- Yusoff Zaharin (switch)
- Ufang Sun (Yufang)
- Horacio Rodffguez (Rodriguez)

There are a large number of these author misspellings on the Website.

10.3 The Papers

The ACL IDs listed in the following tables do not convert cleanly from pdf to txt using PDFbox, producing the noted output. Table 28 shows the failed conversions. Failed means a pdf failed starting the conversion process. Table 29 shows the empty conversions. Empty means that the text extraction produced minimal to no actual text. Table 30 shows conversions with bad output. Gibberish means that the produced text, although appropriate in length, is not human language. This often seems to occur due to strange encodings in the PDF file. As an example, here is the first line of one of these files that produces gibberish text:

a0a2a1a4a3a6a5a8a7a10a9a12a11a14a13a
16a15a17a13a19a18a20a9a22a21a23a13a
16a24a25a1a27a26a28a13a16a15a30a29a
31a11a10a32a34a33a16a15a30a11a34a35a
6a36a37a7a38a1a27a39a40a29a23a29a31a
33a41a13

Table 28: *Problematic Conversions - Failed*

P03-1024	P03-2004	W03-1613	W04-1102	C94-1038

Table 29: *Problematic Conversions - Empty*

C02-1044	C04-1130	E03-1001	E03-1002	E03-1003	E03-1004	E03-1005	E03-1006	E03-1007
E03-1008	E03-1009	E03-1010	E03-1011	E03-1012	E03-1013	E03-1014	E03-1015	E03-1016
E03-1017	E03-1018	E03-1019	E03-1020	E03-1021	E03-1022	E03-1023	E03-1024	E03-1025
E03-1026	E03-1027	E03-1028	E03-1029	E03-1030	E03-1031	E03-1032	E03-1033	E03-1034
E03-1035	E03-1036	E03-1037	E03-1038	E03-1039	E03-1040	E03-1041	E03-1042	E03-1043
E03-1044	E03-1045	E03-1046	E03-1047	E03-1048	E03-1049	E03-1050	E03-1051	E03-1052
E03-1053	E03-1054	E03-1055	E03-1056	E03-1057	E03-1058	E03-1059	E03-1060	E03-1061
E03-1064	E03-1065	E03-1066	E03-1067	E03-1068	E03-1069	E03-1070	E03-1071	E03-1072
E03-1073	E03-1074	E03-1075	E03-1076	E03-1077	E03-1078	E03-1079	E03-1080	E03-1081
E03-1084	E03-1085	E03-1086	E03-1087	E03-1088	E03-2001	E03-2002	E03-2003	E03-2004
E03-2005	E03-2006	E03-2007	E03-2008	E03-2009	E03-2010	E03-2011	E03-2012	E03-2013
E03-2014	E03-2015	E03-2016	E03-2017	E03-3001	E03-3002	E03-3003	E03-3004	E03-3005
E03-3006	E06-1017	E06-2006	H01-1044	H05-1015	J79-1066	J97-3012	N01-1022	N03-2009
N03-2010	N03-2014	N03-5001	N03-5002	N03-5003	N03-5004	N03-5005	N03-5006	N03-5007
N03-5008	N03-5009	N04-1006	N06-3008	P00-1018	P00-1044	P02-1037	P04-1003	P06-4017
P07-2003	W01-1314	W02-0900	W03-1121	W03-1122	W03-1509	W04-0709	W04-0909	W04-1214
W04-2212	W04-2303	W04-3010	W05-1010	W06-0127	W06-1645	W07-0302	W07-0306	W07-0309
C02-1005								

Also, W93-0219 and W93-0220 are problematic. The final pages of W93-0219 are cut off of the PDF, but are then included at the beginning of W93-0220.

Occasionally as well, in the conversion process, pieces are placed out of order. For instance, it was not uncommon to find a few references listed before the heading for the References section was printed. We do not have the actual statistics for this, but it did happen occasionally.

10.4 Other

The following ACL IDs are assigned to the same papers.

- C90-3006/C90-2006
- E99-1029/E99-1042
- C90-3090/C90-3091

The ACL IDs for papers C92-4213 thru C92-4215 link to PDF files that state the papers were ”unavailable at time of print.” Perhaps it should be considered that papers like this now be included in the digital collection after 15 years.

There is a problem with the 2004 Workshops page. The W04-1300’s, W04-1900’s, W04-3000’s, all suffer from an off-by-one kind of error. In each, the website lists the first paper as the Front Matter, and the second as the Introduction/Editorial, when in fact, The Front Matter and Introduction/Editorial are both in the first

Table 30: *Problematic Conversions - Gibberish*

C02-1005	C02-1015	C02-1017	C02-1018	C02-1024	C02-1028	C02-1029	C02-1030	C02-1032
C02-1037	C02-1038	C02-1039	C02-1046	C02-1055	C02-1059	C02-1060	C02-1067	C02-1068
C02-1073	C02-1076	C02-1077	C02-1082	C02-1084	C02-1091	C02-1092	C02-1093	C02-1094
C02-1095	C02-1102	C02-1105	C02-1106	C02-1108	C02-1109	C02-1110	C02-1111	C02-1115
C02-1118	C02-1119	C02-1120	C02-1121	C02-1123	C02-1124	C02-1129	C02-1131	C02-1134
C02-1135	C02-1139	C02-1142	C02-1146	C02-1147	C02-1154	C02-1157	C02-1164	C02-1165
C02-1167	C02-1168	C02-1169	C02-1170	C02-2012	C02-2027	C04-1003	C04-1029	C04-1038
C04-1039	C04-1042	C04-1046	C04-1052	C04-1056	C04-1063	C04-1065	C04-1073	C04-1084
C04-1085	C04-1086	C04-1095	C04-1100	C04-1120	C04-1123	C04-1125	C04-1163	C04-1184
D07-1010	H01-1022	H01-1024	H01-1027	H01-1032	H01-1048	H01-1050	H01-1065	H01-1066
H01-1067	N01-1001	N01-1002	N01-1004	N01-1005	N01-1006	N01-1008	N01-1011	N01-1012
N01-1013	N01-1018	N01-1020	N01-1026	N01-1027	N01-1030	N01-1031	N03-1006	N03-1008
N03-1021	N03-2021	N03-2038	N04-1034	N04-1036	N04-2000	N04-4017	N07-4005	P00-1004
P00-1005	P00-1006	P00-1007	P00-1008	P00-1011	P00-1016	P00-1017	P00-1019	P00-1021
P00-1023	P00-1024	P00-1025	P00-1027	P00-1030	P00-1032	P00-1033	P00-1034	P00-1035
P00-1036	P00-1039	P00-1040	P00-1042	P00-1046	P00-1048	P00-1049	P00-1050	P00-1056
P00-1059	P00-1062	P00-1064	P00-1066	P00-1069	P00-1071	P00-1072	P01-1013	P01-1052
P01-1063	P02-1005	P02-1011	P02-1020	P02-1022	P02-1027	P02-1028	P02-1031	P02-1033
P02-1050	P03-1007	P03-1049	P03-1052	P03-1056	P03-1067	P03-2016	P04-1046	P04-1056
P04-2000	P04-3000	P04-3009	P04-3013	P04-3016	P06-1138	W01-0701	W01-0710	W01-0715
W01-0717	W01-0718	W01-0723	W01-0726	W01-0807	W01-1009	W01-1204	W01-1205	W01-1311
W01-1415	W01-1608	W01-1611	W01-1615	W01-1616	W01-1617	W01-1620	W01-1621	W01-1624
W02-0100	W02-0106	W02-0203	W02-0204	W02-0208	W02-0220	W02-0222	W02-0223	W02-0312
W02-0401	W02-0403	W02-0505	W02-0601	W02-0704	W02-0710	W02-0711	W02-0810	W02-0815
W02-0816	W02-0901	W02-0907	W02-1001	W02-1007	W02-1010	W02-1021	W02-1023	W02-1027
W02-1034	W02-1035	W02-1037	W02-1038	W02-1104	W02-1105	W02-1108	W02-1109	W02-1114
W02-1208	W02-1402	W02-1404	W02-1409	W02-1505	W02-1609	W02-1611	W02-1708	W02-1709
W02-1710	W02-1712	W02-1803	W02-1804	W02-1808	W02-1907	W02-2002	W02-2004	W02-2005
W02-2014	W02-2015	W02-2017	W02-2020	W02-2022	W02-2025	W02-2026	W02-2027	W02-2028
W02-2032	W02-2035	W03-0321	W03-0910	W03-1011	W03-1200	W03-1502	W03-1505	W03-1709
W03-1714	W03-1730	W03-1801	W03-1810	W03-1906	W04-0200	W04-0201	W04-0205	W04-0413
W04-0704	W04-0708	W04-0809	W04-0811	W04-0823	W04-0841	W04-0848	W04-0852	W04-0864
W04-0901	W04-1103	W04-1109	W04-1210	W04-1505	W04-1508	W04-1509	W04-1512	W04-1803
W04-1805	W04-1811	W04-1814	W04-1905	W04-2118	W04-2216	W04-2307	W04-2500	W04-2600
W04-2604	W04-2700	W04-2707	W04-3008	W05-0510	W05-0711	W06-0104	W06-1106	W06-2203
W06-2913	W06-3509							

paper (the one ending in 00). This causes the last two papers in each series, although labeled differently on the website, to point to the same PDF file.

The following Proceedings are absent or not yet classified into the ACL Anthology. We provide this list simply as a reference. We know that some of these are being processed, and that others are not freely available from their source. There may be other reasons that we are not aware of also. But here is the list:

- SIGDAT/EMNLP 2004
- SIGDAT/EMNLP 2001
- SIGDAT/EMNLP 1998
- SIGDAT/WVLC 1994
- COLING 1965 (just the 7 already noted)
- COLING 1971
- COLING 1976
- COLING 1978
- HLT 2002
- MUC 7, 1998
- EACL 2003 Workshops (as noted already), which include:
  - MT and other language technology tools
  - 9th European Workshop on Natural Language Generation
  - 4th International Workshop on Linguistically Interpreted Corpora
  - Language Modeling for Text Entry Methods
  - The Computational Treatment of Anaphora
  - Dialogue Systems: interaction, adaptation and styles of management
  - Computational Linguistics for South Asian Languages
  - Workshop on Finite State Methods in Natural Language Processing
  - Language Technology and the Semantic Web: 3rd Workshop on NLP and XML
  - Natural Language Processing (NLP) for Question-Answering
  - Morphological Processing of Slavic Languages
  - Evaluation Initiatives in Natural Language Processing: are evaluation methods, metrics and resources reusable?
- 2001 Workshops, which include:
  - Automatic Summarization
  - WordNet and Other Lexical Resources: Applications, Extensions and Customizations
  - Arabic Language Processing: Status and Prospects
  - Workshop on MT Evaluation: Hands-On Evaluation
  - Adaptation in Dialog Systems
  - SENSEVAL Workshop