KTH Computer Science
and Communication

Finding seminal scientific publications with graph mining

MARTIN RUNELÖV

Master's Thesis at CSC
Supervisor at CSC: Olov Engwall
Supervisor at SICS: John Ardelius

June 2, 2015

Abstract

We investigate the applicability of network analysis to the problem of finding seminal publications in scientific publishing. In particular, we focus on the network measures betweenness centrality, the so-called backbone graph, and the burstness of citations. The metrics are evaluated using precision-related scores with respect to gold standards based on fellow programmes and manual annotation. Citation counts, PageRank, and random selection are used as baselines. We find that the backbone graph provides us with a way to possibly discover seminal publications with low citation count, and combining betweenness and burstness gives results on par with citation count.

Referat

Användning av grafanalys för att hitta betydelsefulla vetenskapliga artiklar

I detta examensarbete undersöks det huruvida analys av citeringsgrafer kan användas för att finna betydelsefulla vetenskapliga publikationer. Framför allt studeras ”betweenness”-centralitet, den så kallade ”backbone”-grafen samt ”burstness” av citeringar. Dessa mått utvärderas med hjälp av precisionsmått med avseende på guldstandarder baserade på ’fellow’-program samt via manuell annotering. Antal citeringar, PageRank, och slumpmässigt urval används som jämförelse. Resultaten visar att ”backbone”-grafen kan bidra till att eventuellt upptäcka betydelsefulla publikationer med ett lågt antal citeringar samt att en kombination av ”betweenness” och ”burstness” ger resultat i nivå med de man får av att räkna antal citeringar.

Preface

This thesis is a degree project in Computer Science and Communication and concludes the Master's program in Computer Science at the Royal Institute of Technology, KTH. The project was done at the Swedish Intitute of Computer Science, SICS.

Acknowledgements

I would like to thank everyone that provided help and guidance when writing this thesis.

First, I would like to thank my supervisors John Ardelius and Olov Engwall for providing invaluable input throughout the process.

Second, I would like to thank Jonathan Murray for discussing all aspects of the project with me.

Third, I would like to thank my brother, Fredrik Runelöv, for proofreading this thesis.

List of terms

- **Graph** – A set of vertices connected by edges. Graphs are used to refer to concepts such as citation graphs.
- **Network** – A set of nodes and edges. Networks are used to refer to specific datasets modeled as graphs.
- **Seminal** – Something original which has strongly influenced later developments. Influential; groundbreaking.
- **Predecessors** – The predecessors of a node n is the set of nodes with an outgoing directed edge pointing at n.
- **Successors** – The successors of a node n is the set of nodes with an incoming directed edge originating at n.
- **Ancestors** – The ancestors of a node n is the set of all nodes that have a path to n. Ancestors are only defined for directed acyclic graphs.

Contents

Chapter 1

Introduction

This chapter provides an overview of citation analysis and how it can be viewed from a graph perspective, and describes the objective of this thesis.

1.1 Overview

Hundreds of thousands of scientific articles are being published each year. Jinha [1] estimates that about 50 million scientific articles had been published up to 2010. To efficiently search for research in such a large set is a difficult task. There are methods that describe how you can iterate your search phrases and evaluate the results manually as you go, but a more precise solution would be to automatically measure the quality of articles. With better search metrics, researchers could more easily find high quality source material; potentially increasing the quality of research in general. The goal of this thesis is to find methods that can identify publications that have been influential within their subject area. Such publications are likely to interest researchers who want to get a comprehensive overview of a topic or simply find inspiration.

Many popular ranking metrics in use today are based on citation counts in some way. Examples include Hirsch's h-index [2] and Garfield's Impact Factor [3].[1,2] These metrics are often good indicators of popularity but fail to account for many aspects of scientific publishing. Price [4] showed in 1965 that citation counts approximately follow a power law in that the fraction of publications with k citations are proportional to $k^{-\gamma}$ for large k.[3] This can be explained by the fact that publications with

[1] A scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each.

[2] The Impact Factor measures the average number of citations to recent publications in a journal.

[3] γ depends on the data. For example, An, Janssen, and Milios [5] got $\gamma = 1.7$

many citations are more likely to be cited, something which Price called cumulative advantage.[4] Hence metrics based on citation counts can be biased towards popular publications rather than truly influential publications. Newman observed a ”first-mover advantage”-effect in scientific publishing which states that the first papers of a field will receive citations at a significantly higher rate than later published papers regardless of quality [7]. Another issue with citation count based methods is that publishing, co-author, and citation culture differ between research areas, further complicating comparisons between disciplines.

One typical assumption when measuring influence in citation networks is that all references are equal. In reality there are many reasons for citing other publications and influence is unlikely to be equally distributed. Therefore, some researchers have focused on the problem of classifying citations. Zhu et al. [8] did binary classification of references where they were classified as either *influential* or *non-influential* using a large number of features extracted from the full text of publications. Teufel et al. [9] instead used 12 categories, providing even more context.

The next section describes a completely different approach which instead studies the structural properties of references.

1.2 The graph perspective

People have been modeling bibliographic data such as references and co-authorship with graphs for decades, but it has become increasingly popular in recent years with the innovation of graph algorithms, metrics and increased computational power. Hence a number of popular graph concepts already exist within the context of citation analysis:

- *Citation graph* – Directed graph where each vertex is a paper, author, or journal and edges are citations.

- *Co-citation graph* – Undirected weighted graph where each vertex is a paper or author and vertices are connected by an edge if they are co-cited by another paper. More co-citations increase the weight of the edge and it can be used as a measure of similarity (see e.g. figure 1.1).

- *Co-authorship graph* – Undirected graph where each vertex is an author and vertices are connected by an edge if the authors published a paper together.

[4] Also referred to as preferential attachment [6] or ”the rich get richer”

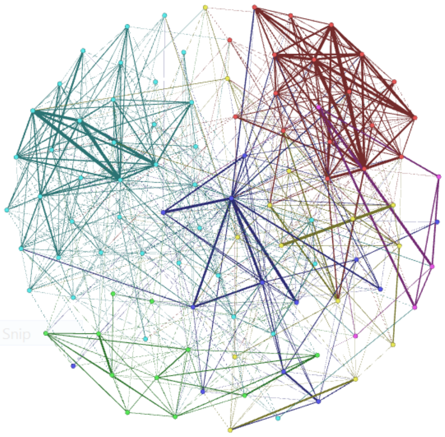


Figure 1.1: Co-citation network of the top 100 most cited publications in a dataset of 1851 papers retrieved from the topic ”citation analysis” in Scopus [10]. The thickness of each edge is proportional to the number of co-citations. Colors indicate separate communities of tightly connected subcomponents detected using the Louvain method [11]

The above graphs can be studied using general graph algorithms and they also have a number of special properties that can be utilized. Citation graphs are, in theory, directed acyclic graphs (DAGs) as citations are made backwards in time.[5] Citation graphs are also dynamic, evolving over time, although one notable difference between citation graphs and other dynamic networks is that citation graphs are static except for additions in the form of new publications with new citations. Another important aspect is that the aforementioned graphs are related and can be expressed as a single heterogeneous network or be used in combination [12, 13, 14].

1.3 Objective

The goal of this thesis is to explore the possibilities of using graph algorithms on citation networks in order to identify seminal publications.

A publication is considered seminal if it has made significant contributions to its subject area, strongly influencing later developments. This is quantified by viewing references as influence indicators and studying the direct and indirect flows of influence that they create in citation networks.

A selection of algorithms for measuring influence in citation networks are implemented and their ability to detect seminal publications is evaluated. The hypothesis is that citation networks contain latent information about the true amount of influence that a publication has had within its

[5] There are some exceptions to the rule; two papers can cite each other before publication, and a paper can reference itself

subject area. Finally, the degree to which the algorithms scale to larger networks is discussed.

1.4 Delimitations

This thesis will focus on the graph perspective and will not include any text analysis of publications. Studying citations using text analysis is a related, but separate, area.

Due to time and resource limitations the evaluation is automated. Given more time and resources human classification by domain experts could be used in order to identify truly seminal publications.

Chapter 2

Theory

This chapter describes the theory used in this thesis. First, the concept of centrality in networks and its applications on citation networks is described. Second, the related concept of influence in citation networks and previous work on the topic is reviewed. Finally, some commonly used evaluation methods are described.

2.1 Centrality measures

In graph theory, *centrality* refers to a measure of importance in a network. Each centrality measure described below has fundamentally different views of importance, leading to different applications in the context of citation networks.

2.1.1 Degree centrality

Degree centrality is the simplest form of centrality and is defined as *the degree of a vertex*. In a directed network this is further divided into *indegree* and *outdegree*. In a citation network, indegree is the number of citations a publication has received and outdegree is the length of its reference list. This kind of centrality is often called *local* centrality, since it is only dependent on the immediate neighbourhood of a vertex, rather than the entire network [15].

2.1.2 Betweenness centrality

The betweenness centrality of a vertex is defined as *the number of shortest paths between all pairs of vertices that the vertex appears on* [16]. For example, vertices that serve as bridges between otherwise disjoint communities in a network will have a high betweenness centrality. In the context of citation networks this can be interpreted as a measure of interdisciplinarity [15]. For example, Leydesdorff found that normalized betweenness centrality is a useful indicator of the interdisciplinarity of journals [17]. The betweenness centrality can be used both on citation networks and on co-citation networks. One drawback of this centrality measure is that it is only defined for undirected networks and therefore does not take the direction of citations into account. This means that a publication can affect its own betweenness centrality using its own references in a regular directed citation network.

2.1.3 Closeness centrality

Closeness centrality measures how well a vertex can reach all other vertices in a network. It is defined as *the average length of the shortest path between a vertex and all other vertices.* It is perhaps the most intuitive of all global measures since it can be interpreted geometrically. The point with the highest closeness centrality is the point which is closest to all other points.

Closeness centrality can be seen as a measure of how well a vertex communicates with the rest of the network. It was originally defined for undirected networks but similar approaches have been suggested to measure closeness centrality in directed networks. Examples include Noh and Rieger's random walk closeness centrality [18] and Tran and Kwon's hierarchial closeness [19]. The original version is, however, applicable to both co-authorship and co-citation networks since they are undirected while one option for citation networks is to make the citation network undirected, an approach used by e.g. Shibata et al. [20].

2.1.4 Eigenvector centrality

Eigenvector centrality, originally defined by Bonacich in 1987 [21], is based on the idea that not all vertices are equal. The basic concept is that the importance of a vertex is affected by the importance of the vertices it is connected to. The score of a vertex is proportional to the sum of the score of its neighbours. This means that a high score can correspond to many connections to unimportant vertices or few connections to important vertices [15]. Eigenvector centrality in its original form is best suited for undirected networks, but there are popular directed alternatives such as Kleinberg's HITS algorithm [22] and Google's PageRank algorithm [23] which is described in detail below.

2.1.4.1 PageRank

PageRank can be modeled using a so-called random surfer model in which a user follows outgoing links at random, gets bored after a while, jumping to a random node. The output is a probability distribution that represents the likelihood that the random surfer will visit a certain page. The PageRank of a node is defined as:

$$PR(v_i) = \frac{1-\alpha}{N} + \alpha \sum_{v_j \in pred(v_i)} \frac{PR(v_j)}{|succ(v_j)|}, \tag{2.1}$$

where N is the number of nodes in the graph, $pred(v_i)$ are the predecessors of v_i, $|succ(v_j)|$ are the number of outgoing edges from v_j and α is the so-called damping factor. PageRank values can be calculated iteratively with PageRank being distributed until it converges.

The damping factor shortens the length of an average walk and distributes PageRank scores across all nodes in the process. Without a damping factor, i.e. a damping factor of 1, all random surfers would eventually end up in sink nodes, giving all other nodes a PageRank of zero. In the original paper by Page et al. [23] a damping factor of 0.85 was suggested.

In citation analysis, PageRank-related methods have received a lot of attention in recent years. Some examples include eigenfactor [24], CiteRank [25], the SCImago Journal Rank (SJR) method [26], P-rank [12] and Chen et al. [27] who suggested that references in scientific publications are collected using shorter paths, corresponding to a damping factor of 0.5. This is partly justified in Chen et al. [27] by their observation that about half of the references in their data have at least one reference pointing to an article that is in the same reference list, i.e a directed triangle where some paper A cites B and C, and B also cites C.

2.2 Influence in citation networks

In the above section many fundamentally different ways to score nodes were introduced. In citation networks a number of different techniques have been used to measure different kinds of influence, typically referred to as centrality, impact, prestige, or authority.[1] In this section previous work related to the task of finding seminal publications is presented.

Leicht et al. [28] examined three different methods for analyzing large-scale networks that evolve over time. Leicht et al. [28] used the expectation-maximization algorithm [29, 30] to classify publications into different temporal profiles in order to, among other things, discover different eras during which certain documents are well-cited. Leich et al. [28] also used Kleinberg's HITS algorithm [22] to calculate authority scores in a citation network and compared the top authorities' average age over time. In

[1] For an overview see, for example, the book ”Measuring scholarly impact” by Ding, Rosseau and Wolfram [15].

a supreme court citation network Leicht et al. [28] observed some sudden drops in the average age, indicating that much younger nodes became authorities in a short period of time. The sudden changes were attributed to specific judges becoming increasingly influential.

Bae, Hwang and Kim [31] proposed an algorithm that extracts a subset of papers that represent an overview of a topic using random walks and link-based similarity measures. Bae et al. outline four types of seminal papers that they want to find, one of which focuses on young seminal papers which haven't had much time to accumulate citations. Gualdi, Medo and Zhang [32] use a similar method to discover seminal papers, proposing an algorithm for DAGs in general and showing that it can be successfully applied to citation networks in order to uncover seminal papers, even ones with low citation counts. One proposal by Gualdi et al. [32] is to normalize rankings using a node's so-called *progeny size* which is the total number of ancestors of a node. Gualdi, Yeung and Zhang [33] proposed a related method for finding the so-called "backbone" of a citation network. The backbone is a pruned version of a citation network where only the most influential references are kept. The result is a descendant chart of publications which can be used to explore chains of influence through a citation network. Gualdi et al. [33] show that the backbone has multiple applications, including identification of seminal papers and other classification tasks. The progeny size can e.g. be used as a quantitative measure of influence since it represents the number of indirectly influenced publications.[2].

Chen et al. [34] propose a model for finding what they call transformative discoveries in which they measure the betweenness centrality of co-citation networks in order to find publications that bridge research areas and create their own, new, research area. This is combined with the concept of burstiness of citation of a reference over time using a burst detection algorithm developed by Kleinberg [35].[3] These two metrics, along with citation counts, are then combined in order to identify seminal publications. The intuition behind Chen et al.'s approach is that transformative discoveries are made when previously disconnected areas are connected by a publication and that the most prominent of these will attract citations in a concentrated manner. They conclude that the normalized geometric mean of betweenness centrality and burstiness, $\sqrt{betweenness \times burstiness}$, identifies seminal publications and partially overcomes the problem where these publications are overshadowed by highly cited publications, such as first-movers. As future work they mention validating the theory on large datasets and further understanding the status of the publications that are highly ranked by these metrics.

[2] For a more in-depth description of the algorithm, see section 2.2.1.

[3] An overview of the burst detection algorithm is given in section 2.2.2

2.2.1 The Backbone algorithm

This section describes the backbone algorithm developed by Gualdi, Yeung and Zhang [33].

Input:
A document citation network

Output:
A "backbone" document citation network, where each document only has one outgoing edge corresponding to the reference which had the biggest impact on the citing document.

We denote the references of a publication as its *parents*, and its incoming citations as its *children* (the corresponding terms in graph theory are *predecessors* and *successors*). The set of parents and children for a publication x are denoted by P_x and C_x respectively. The set $C_x \setminus \{i\}$ is referred to as the *peers* of i rooted in x.

The intuition behind the algorithm is that the children of an influential paper should be similar to each other. Influential publications are assumed to have a more homogeneous descendance than non-influental publications. The *impact* of parent p on a child i is therefore quantified as the sum of all pairwise similarities of the child and its peers (i.e. all other publications that cited p):

$$I_{p \to i} = \sum_{i' \in C_p \setminus \{i\}} s_{ii'}, \tag{2.2}$$

where $s_{ii'}$ is the *similarity* between i and i'.

The definition of the similarity $s_{ii'}$ consists of two parts, $s_{ii'}^{\text{auth}}$ and $s_{ii'}^{\text{read}}$:

$$s_{ii'}^{\text{auth}} = \frac{1}{|P_{i'}|} \sum_{j \in P_i \cap P_{i'}} \frac{1}{|C_j|}, \tag{2.3}$$

$$s_{ii'}^{\text{read}} = \frac{1}{|C_{i'}|} \sum_{j \in C_i \cap C_{i'}} \frac{1}{|P_j|}, \tag{2.4}$$

where $s_{ii'}^{\text{auth}}$ represents the peers (authors) own view of their similarity by looking at their common references. The equation can be viewed as a 2-step random walk from each peer i' to i through their common references. Similarly, $s_{ii'}^{\text{read}}$ represents a reader's view of the similarity between i and i' by looking at publications that cite both peers. Again, a 2-step random walk analogy is used, but through the articles citing both i and i'. $C_i \cap C_{i'}$ is the set of publications that co-cite i and i'.

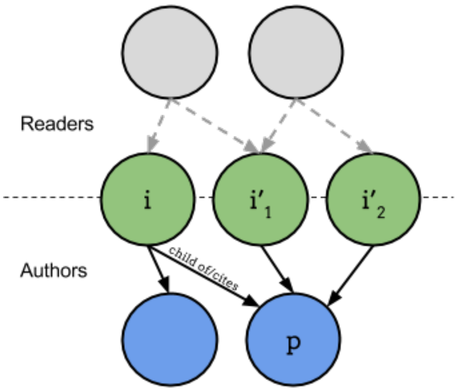


Figure 2.1: Illustration of the reader and author perspectives used to calculate $s_{ii'}^{\text{auth}}$ and $s_{ii'}^{\text{read}}$. Children are the green nodes, parents blue, and publications citing the children are gray.

Finally, $s_{ii'}^{\text{read}}$ and $s_{ii'}^{\text{auth}}$ are linearly combined, giving us the final impact formula:

$$I_{p \to i} = \sum_{i' \in C_p \setminus \{i\}} [f s_{ii'}^{\text{read}} + (1-f) s_{ii'}^{\text{auth}}], \tag{2.5}$$

where f sets the relative weight between $s_{ii'}^{\text{read}}$ and $s_{ii'}^{\text{auth}}$.

The backbone of a citation graph is created by only keeping the most influential references. For a node i the only edge kept is the one to the parent specified by:

$$\arg\max_{p \in P_i} I_{p \to i}. \tag{2.6}$$

2.2.2 Burst detection

This section gives a brief overview of the burst detection algorithm. For additional details see the original paper by Kleinberg [35].

Given a list of reoccuring events with timestamps the burst detection algorithm finds time periods of high intensity. Instead of directly using frequencies the algorithm uses a probabilistic automaton with states that depend on frequency. State transitions occur when there are sudden changes in frequency. There can be a variable number of automaton (bursting) states, representing different levels of intensity. The transitions are modeled using an exponential distribution.

The model is mainly controlled through the following parameters:

- γ – Coefficient for transition costs. A higher γ results in higher transition consts.
- **s** – The base of the exponential distribution used to model event frequencies.
- **states** – The number of automaton states used, with a minimum value of 2 (bursting and non-bursting).

2.3 Evaluation methods

Evaluating influence metrics is a difficult problem in general since influence and originality are subjective concepts. In order to quantitatively measure and compare different ranking algorithms, methods from information retrieval are often used. This involves calculating relevance scores for documents, either by binary classification as either relevant or non-relevant, or through a relevance scale ranging from irrelevant to completely relevant. Relevance scores can be established by examining which publications are connected to prominent people and awards [8, 32] or by letting domain experts classify publications [14, 36].

Additionally, correlation coefficients and how they can be used to compare metrics is described.

2.3.1 Baselines

A baseline algorithm is an algorithm that is used as a point of reference. Baselines allow us to draw conclusions by comparing our results with known approaches. If building upon earlier work, the natural baseline to use is the original work. More general baselines such as degree centrality can also be used. The most basic baseline is one where a completely random model is used. If a suggested model performs worse than randomly selecting elements it is safe to say that it is not performing well.

2.3.2 Precision and recall

In information retrieval with binary classification (e.g. seminal/non-seminal) precision and recall can be used to measure the relevance of a set of retrieved items. It is a common way to evaluate the performance of information retrieval systems where the set of relevant items is known.

Precision is defined as the fraction of retrieved documents that are relevant:

$$\frac{|\{\text{Relevant items}\} \cap \{\text{Retrieved items}\}|}{|\{\text{Retrieved items}\}|}, \tag{2.7}$$

which is equivalent to the fraction of *true positives* in a classification context. One can consider precision at a certain cutoff, where only the top n results are evaluated. This is referred to as *Precision @ n* or *P@n*. This evaluation method can be thought of as an analogy to a real search system in which only the topmost results are interesting.

Recall is defined as the fraction of all relevant documents that are retrieved:

$$\frac{|\{\text{Relevant items}\} \cap \{\text{Retrieved items}\}|}{|\{\text{Relevant items}\}|}. \tag{2.8}$$

Precision and recall are often inversely related. It is always possible to increase recall by increasing the amount of documents retrieved, but this often comes at the cost of precision. One can plot precision and its corresponding recall as the recall goes from 0 to 1 (all items found), giving a view of how the system performs for all data. This is commonly referred to as a precision-recall graph. This is especially useful if a certain recall is desired and one wants to investigate how the precision is affected.

When evaluating rankings of items one can rank all items and then select only a fraction of the top results by using a fixed cut-off score or by selecting a fixed number of items. The precision and recall of the selected items, with respect to a gold standard, can then be calculated and compared with baselines. If the gold standard is a fixed set of publications one can also measure recall, but if the dataset is too big for manual classification one can instead use gold standard criteria, a checklist of desired attributes, in order to manually classify the top results. Since the true number of relevant publications is not known in this case, the recall can only be estimated. This approach, along with other metrics related to precision and recall, is used by Deng et al. [14].

2.3.3 Cumulative gain

Cumulative gain (CG) is a way to measure the overall relevance of a result list. It is a weighted version of *Precision @ n* that assumes a relevance score for each result. The relevance score is often inferred through human judgment where results are graded on a fixed scale ranging from irrelevant to completely relevant, but any relevance score can be used [37]. The formula is simply the sum of all relevance scores for the top p results:

$$\mathrm{CG}_p = \sum_{i=1}^{p} rel_i, \tag{2.9}$$

where rel_i is the assigned relevance of the result at position i.

2.3.3.1 Discounted cumulative gain

Discounted cumulative gain (DCG) is a modified version of CG which takes the position of the results into account. The underlying assumption is that the relative position within the top ranked results is also relevant to the overall score of the results. The DCG formula is as follows:

$$\mathrm{DCG}_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_b(i)}, \tag{2.10}$$

which is identical to the CG formula except that each relevance value is divided by the logarithm of its rank. The discount is not applied at rank 1 since $\log_b(1) = 0$. The base b of the logarithm controls the steepness of the reduction [37].

Another version of DCG places further emphasis on retrieving relevant documents by putting the relevance score on an exponential scale [38]:

$$\mathrm{DCG}_p^{\mathrm{alt}} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_b(i+1)}. \tag{2.11}$$

2.3.4 Correlation coefficients

2.3.4.1 Pearson's correlation coefficient

Pearson's correlation coefficient (PCC,ρ) is a measure of how linearly correlated two variables X and Y are. It is a value in the range $-1 \leq \rho \leq 1$ where a value of 1 means total positive correlation and a value of -1 means total negative correlation. It is defined as the covariance of the two variables divided by the product of their standard deviations:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{2.12}$$

In citation analysis the PCC can be used to measure the correlation between different scoring metrics. If, for example, a proposed scoring metric is strongly positively correlated with a naive metric such as citation count then it might not be a valuable result since it is trivially obtained. If two metrics are strongly correlated one can simply choose the simplest and most effective metric [12, 20, 34]. However, care must be taken when studying PCC's. Pearson himself warned people against comparing two metrics with common factors [39], and as shown by West et al. in *"Big Macs and Eigenfactor scores: Don't let correlation coefficients fool you"* [40] it is important to tread lightly when studying the magnitude of correlation coefficients. In citation analysis there are often common factors to consider and correlation coefficients are often used as a complement to other evaluation methods.

2.3.4.2 Spearman's rank correlation coefficient

Spearman's rank correlation coefficient is defined as the Pearson correlation coefficient between the ranked variables. It is calculated by ranking variables based on their score and using the difference in ranking orders instead of the difference in values as input to the Pearson correlation coefficient[41].

2.3.5 Logistic regression

In regression, the relationship between dependent and independent variables is investigated. Logistic regression, also called *logit*, is a statistical model for analyzing the effects of multiple independent variables on a dependent, binary, variable. The logistic model predicts class probabilities based on some estimation method and can therefore be used as a classifier. One commonly used method is Maximum Likelihood Estimation (MLE) in which the coefficients which maximizes the probability of observing the input data is chosen [42]. If the estimated probabilities are used to classify instances one can evaluate the performance of the classification using precision and recall.

2.4 Summary

Many possible ways of measuring influence in citation networks have been presented. The effectiveness of some of the suggested metrics will be explored on different datasets. The backbone algorithm will be implemented and its applicability to the task of finding seminal papers will be explored. This involves building the backbone of the citation graph and calculating each node's progeny size.

The following metrics will be explored:

- Citation count (indegree)
- Backbone progeny size
- PageRank
- Betweenness centrality in co-citation networks
- $\sqrt{betweenness \times burstiness}$
- $\sqrt[3]{betweenness \times burstiness \times indegree}$

In addition, logistic regression on the above metrics will be explored. The hypothesis is that the metrics can be used as indicators of seminality and that none of them are trivially correlated.

Chapter 3

Method

This chapter describes the method in more detail. The main steps are as follows:

1. Collect citation data and gold standards
2. Build citation graph, co-citation graph and backbone graph
3. Score the publications using the different metrics
4. Rank the publications according to their score (sort score in descending order)
5. Evaluation of the ranking orders

The metrics that will be evaluated are betweenness centrality in co-citation networks, both as a standalone metric and combined with burstiness and citation count using their normalized geometric mean, and the backbone progeny size. A weighted version of the progeny size, which places further emphasis on having more direct progeny, will also be tested. The weight used is $1/2^{\texttt{depth}}$ where depth is the distance from the root node. This is analogous to how the damping factor, α, in PageRank controls the average length of citation chains.

In addition, citation count and PageRank will be used as baselines and logistic regression is used in an attempt to combine the suggested metrics. The logistic regression uses citation count, the backbone progeny size, and all metrics that include the betweenness centrality as independent variables, and a gold standard (seminal/non-seminal) as the binary dependent variable. The metrics, and when applicable their respective parameters, are summarized below:

- Citation count (indegree)

- Backbone progeny size, the relative weight between $s_{ii'}^{\text{read}}$ and $s_{ii'}^{\text{auth}}$, f, is set to 0.5, the default value
- PageRank, damping factor, $\alpha = 0.5$ in accordance with the findings of Chen et al. [27]
- Betweenness centrality in co-citation networks
- (Burstiness, $\gamma = 1$, $s = 2$, states $= 2$, see section 2.2.2)
- $\sqrt{betweenness \times burstiness}$
- $\sqrt[3]{betweenness \times burstiness \times indegree}$
- Logistic regression using Maximum Likelihood Estimation

Burstiness is written within parentheses since it is only used in combination with betweenness and never directly.

3.1 Data sources

In order to evaluate the algorithms, a dataset containing citation information and certain metadata such as titles, authors, and publication dates is needed. There are many possible resources to choose from, each with their own strengths and weaknesses. The main difference between different datasets is their size and the degree to which publications have been preprocessed. Some only provide full text versions of publications, leaving metadata extraction to the user, while some have minimal or flawed extraction methods. In general it is hard to find a balance between size and quality.

In order to systematically evaluate and compare different data sources a decision matrix was used. Five attributes were chosen and their relative importance was calculated. The attributes are: size, quality of metadata, completeness (level to which it contains cited works), amount of duplicates, and time frame. Some of the attributes are correlated, and time frame received a weight of 0, but the score is still an useful indicator of how well the datasets fit the needs of this thesis. See Appendix A for a full description of the attributes and the score calculations. The final scores are shown in Table 3.1.

Table 3.1: The scores of the datasets with respect to size, metadata quality, completeness and duplication.

Data source	**#Publications**	**#References**	**Score**
The ACL Network Anthology Corpus (AAN) [43]	18 375	110 975	160
The American Physical Society Corpus (APS) [44]	~530 000	~6 000 000	140
KDD Cup 2003 Datasets (hep-th, from arXiv) [45]	~29 000	550 388	130
Scopus [10]	N/A	N/A	130
arXiv [46]	N/A	N/A	80
CORE [47] (2014 dump)	369 696	N/A	40

Scopus and arXiv are missing size information since they are not precompiled datasets. arXiv is an e-print archive and Scopus is a citation database from which you can download up to 2000 search results at a time (with a subscription).

The top two data sources, AAN and APS, will be used to evaluate the algorithms presented in this thesis. Special focus will be given to APS since its superior size makes it more fitting as an example of real-world large-scale publication datasets.

In order to evaluate the ranking orders we make use of compiled lists of prominent authors. Both APS and AAN have fellowship programs that elect fellows who have made significant scientific contributions within their field, meaning that membership can be used as an influence indicator [8, 48, 49]. In order to utilize these lists it is assumed that, for each author in the list, all of their publications in the dataset are seminal. This approach leads to many Type I errors (false positives) since it is often not all of an author's work that is considered seminal, and there will also be Type II errors (false negatives) since many seminal publications will have been written by people that are not present in the list. However, the benefit of this approach is that we are able to select a large subset of publications that are connected to human judgment on seminality. Drawbacks and alternatives are further discussed in section 5.2.2.2.

In the following sections the two datasets, along with a description of their respective fellow programs, is described.

3.1.1 APS

This dataset is maintained by the American Physical Society (APS) and consists of about 530 000 publications and 6 million citations between these publications. Only internal citations, where both articles are in the dataset, are present in the data. The earliest article in the dataset was published in 1893.

3.1.1.1 APS Fellows

APS elects fellows amongst its members each year. The fellowship program was founded in 1980. They give the following description of the selection process:

> *Any active APS member is eligible for nomination and election to Fellowship. The criterion for election is exceptional contributions to the physics enterprise; e.g., outstanding physics research, important applications of physics, leadership in or service to physics, or significant contributions to physics education. Fellowship is a distinct honor signifying recognition by one's professional peers* [50].

3.1.2 AAN

This dataset is maintained by the Association of Computational Linguistics (ACL) and consists of 18375 publications and 110975 citations between these publications. Only internal citations, where both articles are in the dataset, are present in the data. The earliest article in the dataset was published in 1965.

3.1.2.1 ACL Fellows

ACL elects fellows amongst its members each year. The fellowship program was founded in 2011. They describe the program as follows:

> *[...] ACL Fellows program, which recognizes ACL members whose contributions to the field have been most extraordinary. To be named a Fellow, a candidate must have been a member of the ACL for the past three consecutive years and be nominated by a current ACL member* [51].

3.2 Preprocessing

Real-world data such as the datasets used in this thesis often contain errors and inconsistencies, and is therefore sometimes referred to as *dirty data* since some cleanup is required. A few preprocessing steps that was introduced in order to clean up the data used are described below.

3.2.1 Self-citations

Self-citations can be considered on different levels of granularity. For example, when studying inter-journal citations one might want to remove or group all intra-journal citations. When studying how authors interact one might want to ignore an author's self-citations where they cite their own previous work.

In this thesis only the most direct form of self-citation is considered, where a publication cites itself. Such a citation should not contribute to any influence rankings and is therefore removed. 46 self-citations were removed from the AAN graph. There were no self-citations in the APS data.

3.2.2 Forward citations

When exact timestamps are available for most or all of the publications one can compare the date of the citing publication to that of the cited publication. Citations to publications with a later publication date than the citer are here referred to as *forward citations*. Only allowing references that point backwards in time makes sure that the entire graph is directed and acyclic (a DAG).

The following pruning based on date metadata was performed on the datasets:

- **APS** – 4130/531478 (0.78%) publications were missing date information. 18048/5994180 (0.3%) citations were forward citations. All of these were removed. The resulting network is a DAG.
- **AAN** – Only publication year available. No publications were missing date information. There were no forward citations. The network is not a DAG.

3.2.3 Isolated nodes

Isolated nodes, also referred to as isolates, are nodes that have no incoming or outgoing edges. Since such nodes do not contribute to any rating and can not receive any rating based on edges they are removed from the networks. This preprocessing step is performed last since other steps might have removed edges and created new isolates. 217 isolates were removed from the AAN dataset. The were no isolates in the APS data.

3.3 Implementation

All metrics were calculated using Python. The backbone algorithm was implemented using the graph library *NetworkX* [52] and the backbone graph was computed on a server provided by SICS. The construction of the citation and co-citation graphs, including preprocessing, was implemented. NetworkX proved to be too slow for certain calculations which led to Indegree, PageRank and Betweenness centrality being calculated using the graph library *graph-tool* [53] which has algorithm implementations written in C++. Constructing the co-citation graph and calculating its betweenness centralities for the APS data was the most time-consuming task since the graph has over 33 million edges.

The logistic regression is done using the data analysis package *pandas* [54] and the statistical modeling package *Statsmodels* [55]. The *Sci*^{2} [56] tool was used for the burstiness calculations with a burst length unit of *Year*, meaning that month and day information was discarded.

Betweenness, burstiness and indegree were all normalized before being combined. The metrics were also normalized before use in the logistic regression.

Matching author names in the fellow lists to author names in the datasets is done by assuming that people with the same name are the same person. If one name contains additional information, e.g. when comparing ”Martin Runelöv” to ”Martin J. Runelöv”, the missing information is assumed to be correct. Most of the data used has both first names and last names, and in many instances initials for middle names, minimizing the effect of this assumption. The AAN dataset contained many small errors, with one or a few incorrect or missing letters due to erroneous parsing. Therefore, a string similarity check was introduced using the built-in Python class `SequenceMatcher` [57] which compares pairs of sequences and outputs a number between 0 and 1 where 1 means that the sequences are identical. Last names were required to have a match of at least 0.8 and first names a match of at least 0.9. Initials in middle names were still required to fully match. The percentages were set using trial and error with manual checking of the resulting matches.

3.4 Evaluation

Both Pearson’s and Spearman’s correlation coefficient will be calculated for all metrics, pairwise. Calculating Pearson’s correlation coefficient for normalized values of the metrics allows us to study how the magnitudes of the scores differ between metrics, while Spearman’s correlation coefficient, analogous to how precision only takes order into account, allows us to quantify the differences in ranking order between metrics.

The metrics are evaluated using precision, recall and DCG scores with respect to the gold standards based on the fellowship programs. When calculating the DCG the relevance score is binary, with a value of 1 if the publication is in the gold standard and 0 otherwise. Since the relevance score is binary, the DCG$^{\text{alt}}$ formula is used in order to more sufficiently highlight differences in ordering. For each dataset, two ranking variants are used: One where the entire dataset is considered, and one where only publications published after a certain year was considered. For the APS dataset the year 1980 was chosen since the fellow program started in 1980. This removes publications that might be seminal but that often are too old to have been written by fellows. For the AAN dataset the year 2000 was chosen. The fellow program was started in 2011, but since too little time has passed since then the year 2000 was chosen in order to retain a sufficient amount of data and fellow publications.

In order to test the stability of the metrics an additional experiment is performed for the backbone progeny size as well as the baselines Indegree and PageRank on the APS dataset. The data is randomly split into two equally sized parts ten times and the mean precision and standard deviations for all of these subsets is calculated. This allows us to check how the metrics perform on the data in general, making sure that the results are not overly sensitive to changes in the underlying data. The betweenness-related metrics was excluded from this test due to lack of time. A co-citation graph needs to be built for each new random half of the dataset, and this, along betweenness calculations on the co-citation graph, is very time-consuming.

Finally, in order to overcome possible issues with the gold standards used, a manual annotation step is performed where top scoring publications are studied in more detail. Similarly to how Gualdi, Medo and Zhang [32] evaluate their results we study how the different rankings relate to each other and check if any authors have made major contributions to an area related to the article according to Wikipedia [58]. This is done by examining the Wikipedia article of the authors of a publication, noting specific mentions of contributions, nominations and awards related to the publication or its field.

Chapter 4

Results

In this chapter three kinds of results are presented:

- Pearson's and Spearman's correlation coefficients for all metrics, pairwise.
- *Precision @ n* graphs showing the fraction of fellow publications among top ranked publications
- Manually annotated lists of the top 20 results for selected metrics

The precision-related graphs come in two variants: one for the whole dataset and one with an age filter that only considers publications published after a specified year. The amount of fellow publications (true positives) and total amount of publications available in each case are summarized in Table 4.1.

The results from the stability experiment where the dataset is randomly split into halves is presented for the APS dataset.
The metric $\sqrt{betweenness \times burstiness}$ is referred to as **g** (geometric mean), $\sqrt[3]{betweenness \times burstiness \times indegree}$ is referred to as **g2**, and the backbone progeny size is referred to as **BPS**.

Table 4.1: The amount of fellow articles (true positives) available in each dataset. The cutoff column specifies the earliest allowed publication year.

Dataset	**Cutoff**	**Fellow articles**	**Total**	**% Fellow articles**
APS	–	160946	527130	30.5%
APS	1980	139755	427735	32.7%
AAN	–	1603	18158	8.8%
AAN	2000	1197	13589	8.8%

4.1 Correlation coefficients

Pearson's and Spearman's correlation coefficients between all pairs of the metrics used are presented below. A value of 0 means that there is no linear correlation and a value of 1 means that the metrics are completely positively correlated. First, Pearson's correlation is presented, which measures the linear correlation between the *normalized values* of each metric. Second, Spearman's correlation coefficient is presented, which measures the linear correlation between the *ranking orders* of each metric. We are especially interested in each metric's correlation with indegree, and the correlation between burstiness and betweenness, both individually and when combined in **g** and **g2**.

4.1.1 Pearson's correlation coefficient

The Pearson correlation coefficients on the APS data is presented in Table 4.2. It shows that **BPS** has very low correlation with all other metrics, including Indegree. **g** and **g2** both have slightly higher correlations with Betweenness compared to Burstiness and Indegree, indicating that the Betweenness values had the biggest impact on the geometric means. However, Betweenness and Burstiness show relatively low correlation, indicating that we get non-trivial results when combining them. The high correlation between **g** and **g2** is expected due to their low correlation with Indegree and their similar definitions.

The Pearson correlation coefficients on the AAN data is presented in Table 4.3. It shows trends similar to those described above for the APS data, with some exceptions, and with higher correlations overall. **g** and **g2** show comparable correlations with Indegree, Betweenness, and Burstiness, unlike APS where Betweenness stood out. The increased correlation with Indegree can be explained by the increased correlation between Betweenness and Indegree. Burstiness and Betweenness show higher, but still relatively low, correlation.

Table 4.2: Pairwise Pearson correlation coefficient for the metrics on the **APS** data, rounded to three decimals. Low values are marked with an asterisk (*). All correlations are positive.

	Betweenness	**Burst weight**	**BPS**	**g**	**g2**	**PageRank**
Indegree	0.452	0.745	0.125*	0.511	0.567	0.817
Betweenness		0.255	0.101*	0.754	0.752	0.466
Burst weight			0.128*	0.503	0.518	0.660
BPS				0.118*	0.110*	0.287
g					0.980	0.518
g2						0.541

Table 4.3: Pairwise Pearson correlation coefficient for the metrics on the **AAN** data, rounded to three decimals. All correlations are positive.

	Betweenness	**Burst weight**	**BPS**	**g**	**g2**	**PageRank**
Indegree	0.859	0.638	0.255	0.821	0.884	0.780
Betweenness		0.469	0.236	0.800	0.830	0.733
Burst weight			0.260	0.832	0.809	0.692
BPS				0.288	0.285	0.509
g					0.986	0.807
g2						0.808

4.1.2 Spearman's correlation coefficient

The Spearman correlation coefficients on the APS data is presented in Table 4.4. The coefficients are similar to the corresponding Pearson coefficients (Table 4.2), with some exceptions; **BPS**, **g**, **g2** and PageRank all show slightly higher correlation with Indegree, and **g** and **g2** are very strongly correlated, indicating that they produce very similar rankings.

The Spearman correlation coefficients on the AAN data is presented in Table 4.5. The very high correlation between Burstiness and the geometric means indicates that ranking based on Burstiness is almost identical to ranking based on **g** or **g2**. A likely explanation for this is that the Burstiness was 0 for many publications, leaving them in their original positions, unranked. Similar to the difference in Pearson coefficients between the datasets, Betweenness is more correlated with Indegree compared to APS.

Table 4.4: Pairwise Spearman correlation coefficient for the metrics on the **APS** data, rounded to three decimals. Low values are marked with an asterisk (*). All correlations are positive.

	Betweenness	**Burst weight**	**BPS**	**g**	**g2**	**PageRank**
Indegree	0.305	0.718	0.229	0.328	0.328	0.907
Betweenness		0.257	0.075*	0.726	0.726	0.285
Burst weight			0.242	0.455	0.455	0.660
BPS				0.122*	0.122*	0.344
g					0.999	0.311
g2						0.311

Table 4.5: Pairwise Spearman correlation coefficient for the metrics on the **AAN** data, rounded to three decimals. All correlations are positive.

	Betweenness	**Burst weight**	**BPS**	**g**	**g2**	**PageRank**
Indegree	0.918	0.507	0.413	0.508	0.509	0.925
Betweenness		0.506	0.252	0.509	0.509	0.813
Burst weight			0.285	0.998	0.998	0.480
BPS				0.286	0.286	0.546
g					0.999	0.481
g2						0.481

4.2 Finding fellows

This section presents the results from the experiments where the metrics were evaluated by how well they could identify publications written by fellows. First, the results for the APS data is presented, including results from the stability experiment where the dataset is randomly split into halves. Second, the results for the AAN data is presented.

4.2.1 APS

Figure 4.1 shows the precision for lists of length 10-500. Logit shows the best performance and is the only metric to reach a precision above 0.6 (60%). Indegree has the highest precision among the rest of the metrics for lists above length 60. **BPS** show very bad performance, below Random retrieval. The weighted version of the progeny size performs much better, on par with PageRank and betweenness, but all three are only marginally better than Random Retrieval. Figure 4.2 shows the precision for lists of length 10-500 with results filtered to only include publications published in or after the year 1980. All metrics except Logit show improvement, especially **BPS**, and they all outperform Random retrieval significantly. **g** and **g2** perform marginally better than the other metrics overall, except for lists of length 10-20 where PageRank has very high precision.

Figures 4.3 (a)-(f) shows the results from the stability experiment. All figures show that the variance is large for short lists and then decreases significantly, indicating that some variations in precision for short lists is to be expected. **BPS** for the entire dataset (a) is the only metric that deviates significantly from the mean curve. PageRank with a 1980 cutoff (f) shows that it usually performs well for short lists, and both Indegree plots, (c) and (d), show that it usually has low precision for very short lists but then quickly improves.

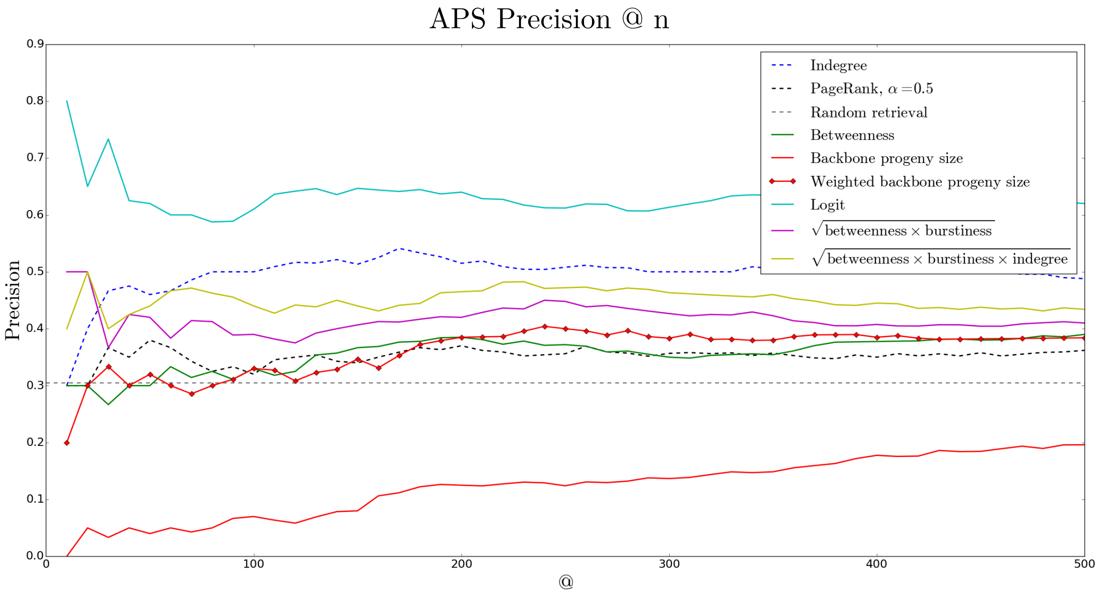


Figure 4.1: The fraction of fellow publications among top ranked publications.

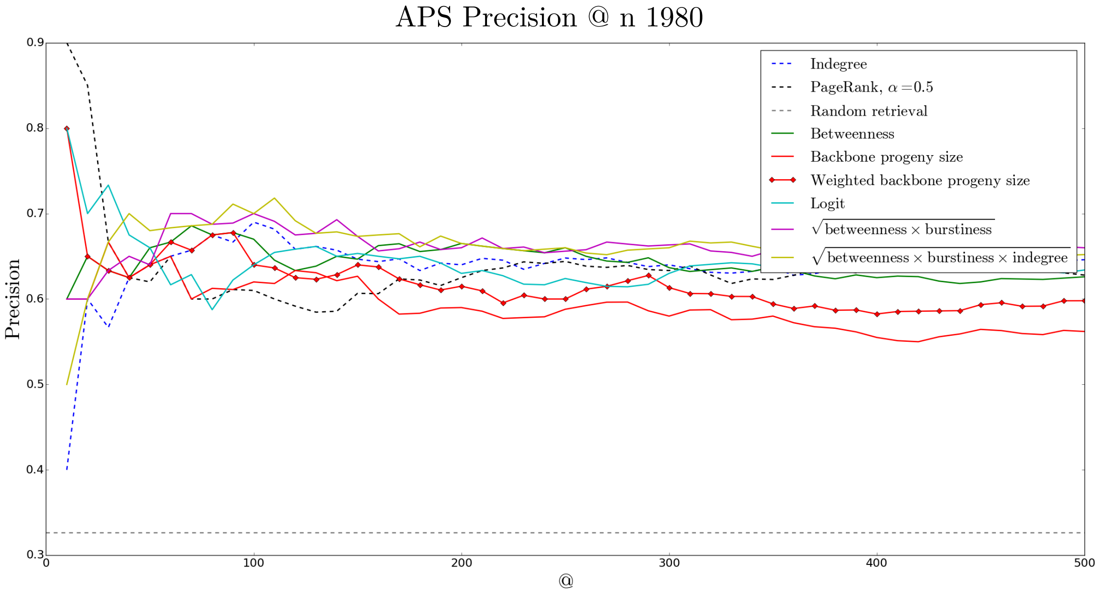
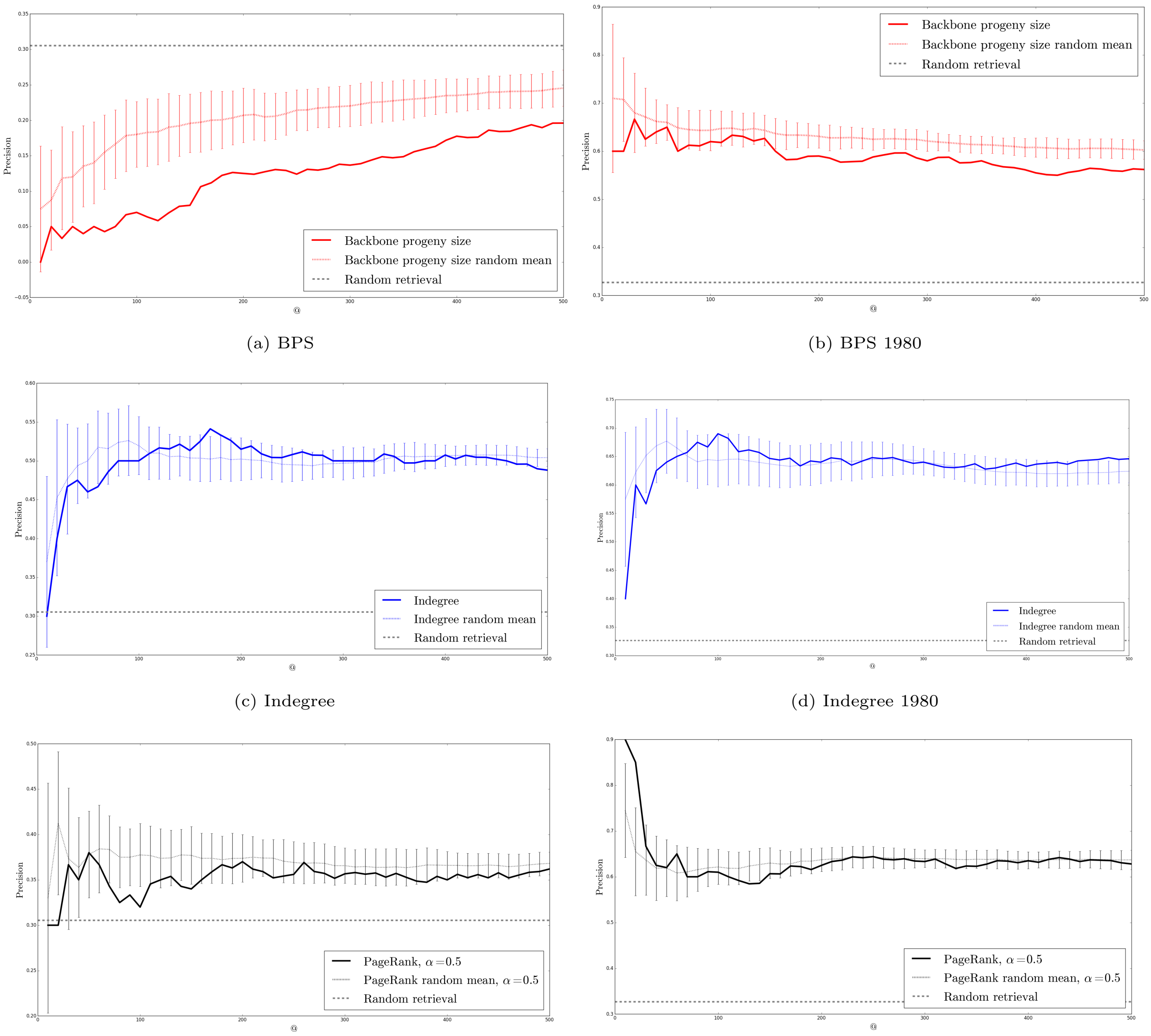


Figure 4.2: The fraction of fellow publications among top ranked publications with results filtered to only include publications published in or after the year 1980.

4.2.1.1 Stability experiment



(a) BPS

(b) BPS 1980

(c) Indegree

(d) Indegree 1980

(e) PageRank

(f) PageRank 1980

Figure 4.3: (a)-(f). The fraction of fellow publications among top ranked publications where thick lines were calculated using the entire dataset and the dotted lines show the mean value for 20 randomized halves of the dataset (10x2 halves). The vertical lines from the dotted line show one standard deviation from the mean. The left column includes the entire APS dataset and the right column is with results filtered to only include publications published in or after the year 1980.

4.2.2 AAN

Figure 4.4 shows the precision for lists of length 10-500. With the exception of fluctuations for small lists, all metrics except **BPS** show similar performance. The weighted variant of **BPS** performs much worse than its unweighted counterpart. Figure 4.5 shows the precision for lists of length 10-500 with results filtered to only include publications published in or after the year 2000. All metrics perform marginally worse than for the whole dataset, except for the unweighted BPS which shows improvement. The weighted BPS performs slightly worse than without a cutoff year. Betweenness still has the highest precision for smaller lists.

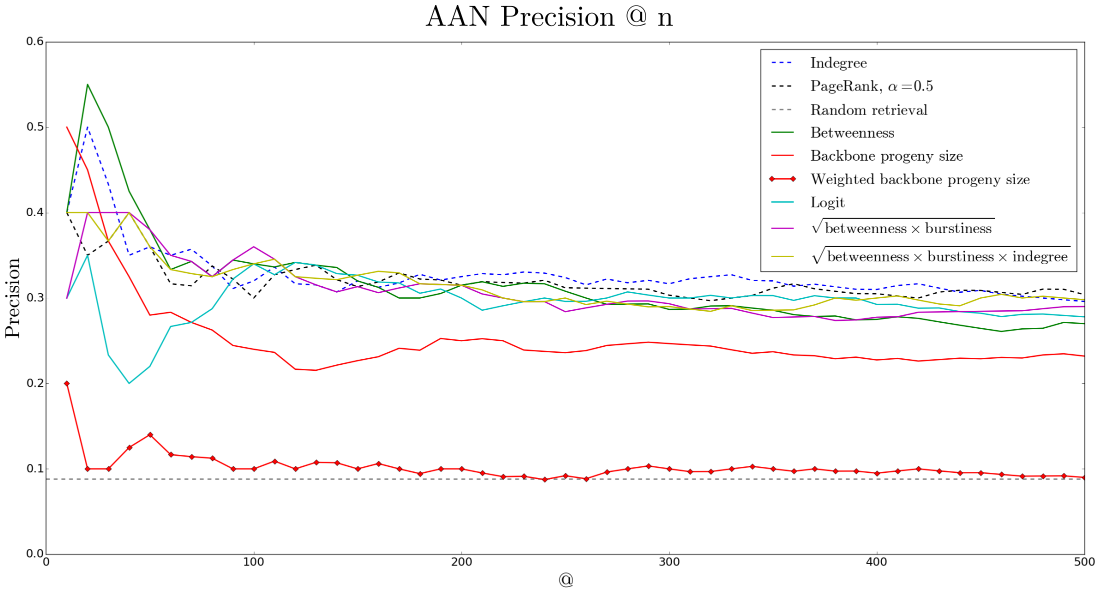


Figure 4.4: The fraction of fellow publications among top ranked publications in the **AAN** dataset.

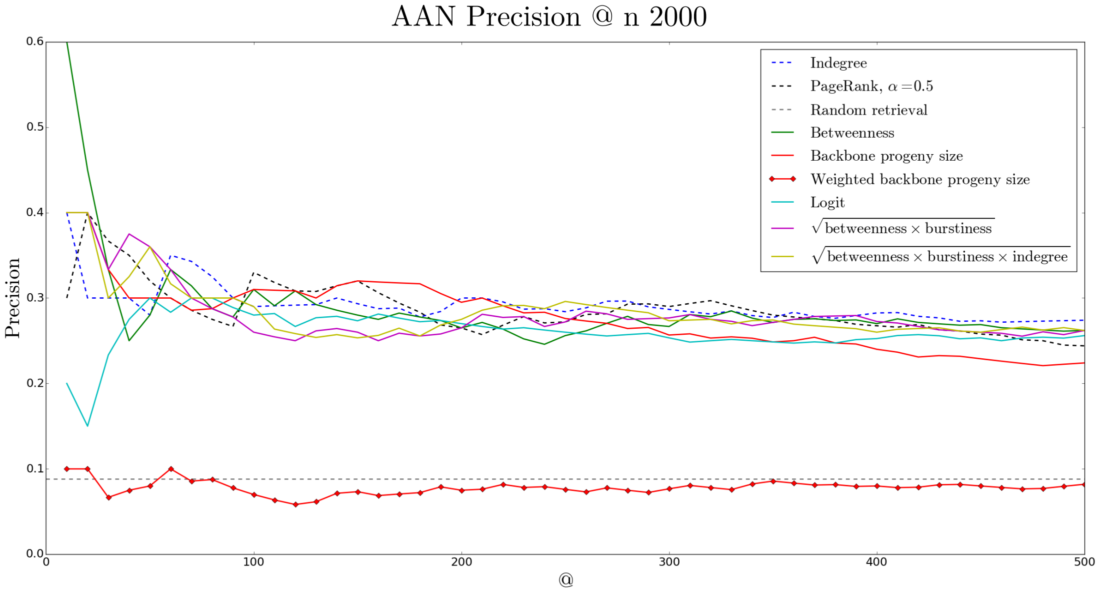


Figure 4.5: The fraction of fellow publications among top ranked publications in the **AAN** dataset with results filtered to only include publications published in or after the year 2000.

4.3 Top results in detail

This section presents the manually annotated lists of top scoring publications in the APS dataset.

The lists were created by extracting the top 20 results of the backbone progeny size for all publications and $\sqrt{betweenness \times burstiness}$ for publications published in or after **1980**. The backbone progeny size performed poorly compared to other metrics (see Figure 4.1), while the geometric mean was among the best (see Figure 4.2).

Each result (row) has one of four colors: **red** if the paper has been deemed non-seminal, **green** if the paper was written by a fellow, **blue** if one of the authors of the paper has made major contributions to a subject related to the paper according to Wikipedia, and **white**, the default color, if no remarks are made. A summary of the colors used is shown in Table 4.6.

The list for the backbone progeny size is presented in Table 4.7. We see that all publications have very early publication years, starting at 1928 for the top result and then gradually decreasing. All results also have low citation counts and show vast differences in ranking order compared to **g**, **g2** and Indegree. #10 has been deemed non-seminal since it is minutes of a meeting. An interesting observation is that the top four results are all written by the same author, J.C. Slater.

The list for $\sqrt{betweenness \times burstiness}$ is presented in Table 4.8. The precision is 60% and all publications are well-cited. The corresponding ranking orders for **g2** and Indegree are low, indicating that they are correlated. Only one additional publication is deemed seminal by the Wikipedia criterion.

Table 4.6: Color definitions

Color	**Definition**
🟩	Fellow
🟦	Wikipedia mention
🟥	Non-seminal
	Default

Table 4.7: Top 20 results for the **Backbone progeny size** metric.

BPS-rank	g-rank	g2-rank	Indegree-rank	Indegree	Title	Authors	Year	Fellow
1	170821	170821	149626	10	The Self Consistent Field and the Structure of Atoms	J. C. Slater	1928	No
2	317	479	3332	133	The Theory of Complex Spectra	J. C. Slater	1929	No
3	47500	47446	260740	5	Central Fields and Rydberg Formulas in Wave Mechanics	J. C. Slater	1928	No
4	166618	166618	26284	40	The Normal State of Helium	J. C. Slater	1928	No
5	7026	9839	84277	17	The Role of Quadrupole Forces in Van Der Waals Attractions	Henry Margenau	1931	No
6	195931	195931	100242	15	Van der Waals Potential in Helium	Henry Margenau	1939	No
7	179044	179044	16263	54	The Theoretical Constitution of Metallic Beryllium	Conyers Herring; A. G. Hill	1940	No
8	179044	179044	16263	54	Scattering of Electrons in Ionized Gases	Irving Langmuir	1925	No
9	1814	2231	11481	66	Oscillations in Ionized Gases	Lewi Tonks; Irving Langmuir	1929	No
10	84066	84066	392032	2	Minutes of the Spring Meeting of the New England Section at Brown University [...]		1955	No
11	251349	251349	81739	18	Magnetic Field Dependence of Ultrasonic Attenuation in Metals at Low Temperatures	Sergio Rodriguez	1958	No
12	244882	244882	210761	7	Elastic Scattering of Yukawa Particles. I	Otto Laporte	1938	No
13	242108	242108	98382	15	The Production of Soft Secondaries by Mesotrons	J. R. Oppenheimer; H. Snyder; R. Serber	1940	No
14	200343	200343	29992	37	Burst Production by Mesotrons	R. F. Christy; S. Kusaka	1941	No
15	238466	238466	324825	3	Meson Lifetime and Radioactive Decay	S. Rozental	1941	No
16	213250	213250	345128	2	Meson Theory and the Magnetic Moments of Protons and Neutrons	H. Fröhlich	1942	No
17	213249	213249	214104	6	Meson Theory of the Magnetic Moment of Proton and Neutron	J. M. Jauch	1943	No
18	442987	442987	223839	6	A Convergent Expression for the Magnetic Moment of the Neutron	D. Rivier; E. C. G. Stuecklberg	1948	No
19	28418	27753	88597	17	Magnetic Oscillations of Ultrasonic Attenuation in a Copper Crystal at Low Temperatures	R. W. Morse; J. D. Gavenda	1959	No
20	554	740	4012	120	Cancellation of Kinetic and Potential Energy in Atoms Molecules and Solids	Morrel H. Cohen; V. Heine	1961	Yes

Table 4.8: Top 20 results for $\sqrt{betweenness \times burstiness}$ for publications published in or after **1980**.

g-rank	g2-rank	BPS-rank	Indegree-rank	Indegree	Title	Authors	Year	Fellow
1	3	37292	54	1100	Review of Particle Properties	K. Hagiwara et. al. (Particle Data Group)	2002	Yes
2	1	5799	8	2815	From ultrasoft pseudopotentials to the projector augmented-wave method	G. Kresse; D. Joubert	1999	No
3	6	1654	20	1550	The electronic properties of graphene	A. H. Castro Neto; F. Guinea; N. M. R. Peres et al.	2009	No
4	7	1369	18	1591	Disordered electronic systems	Patrick A. Lee;T. V. Ramakrishnan	1985	Yes
5	8	51820	32	1344	Spintronics: Fundamentals and applications	Igor uti;Jaroslav Fabian;S. Das Sarma	2004	No
6	9	2052	19	1556	Electronic properties of two-dimensional systems	Tsuneya Ando;Alan B. Fowler;Frank Stern	1982	Yes
7	10	2347	23	1525	Large Mass Hierarchy from a Small Extra Dimension	Lisa Randall;Raman Sundrum	1999	Yes
8	13	3523	35	1290	Theory of Bose-Einstein condensation in trapped gases	F. Dalfovo; S. Giorgini; L. P. Pitaevskii; S. Stringari	1999	No
9	14	35330	28	1417	Bose-Einstein Condensation in a Gas of Sodium Atoms	K. B. Davis et al.	1995	Yes
10	15	7966	25	1483	Vortices in high-temperature superconductors	G. Blatter et al.	1994	Yes
11	17	152506	49	1122	iColloquium/i: Topological insulators	M. Z. Hasan;C. L. Kane	2010	Yes
12	18	9319	45	1146	Unified Approach for Molecular Dynamics and Density-Functional Theory	R. Car;M. Parrinello	1985	Yes
13	34	12505	247	522	Review of Particle Properties		1992	No
14	12	6039	9	2453	Efficient pseudopotentials for plane-wave calculations	N. Troullier;Jos Luriaas Martins	1991	No
15	23	5767	67	994	Pseudopotentials that work: From H to Pu	G. B. Bachelet;D. R. Hamann;M. Schlter	1982	No
16	20	12246	33	1328	An Alternative to Compactification	Lisa Randall;Raman Sundrum	1999	Yes
17	38	11446	167	634	Review of Particle Properties	L. Montanet;K. Gieselmann; R. M. Barnett; C. Grab et al. (Particle Data Group)	1994	Yes
18	22	10355	31	1345	Quantum computation with quantum dots	Daniel Loss;David P. DiVincenzo	1998	Yes
19	27	4379	43	1180	Metal-insulator transitions	M. Imada; A. Fujimori; Y.Tokura	1998	No
20	30	9405	70	979	Evidence of Bose-Einstein Condensation in an Atomic Gas with Attractive Interactions	C. C. Bradley;C. A. Sackett;J. J. Tollett;R. G. Hulet	1995	Yes

Chapter 5

Conclusions and discussion

This chapter is divided into two parts. First, conclusions made from the experiments are summarized. Second, the metrics and the results are analyzed in more detail and time complexities, possible sources of error, ethics and future work are discussed.

5.1 Conclusions

Several different ways of modeling influence have been suggested and their applicability to the task of finding seminal publications have been evaluated. Betweenness centrality and the geometric mean of Betweenness centrality and Burstiness shows promise, with performance often on par with citation count. However, Betweenness was highly correlated with Indegree in the AAN data, suggesting that it might not always provide additional information and that indegree might be preferable due to the complexity of the Betweenness calculations.

The backbone progeny size identified interesting publications with very low citation counts. Considering that most metrics in use today are somehow based on citation counts the backbone graph could help us identify seminal publications that would otherwise go unnoticed. However, the backbone progeny size's bias towards older publications has to be considered before use. The weighted variant of the backbone progeny size seemed to counteract the age bias in the APS dataset, but in the AAN dataset, which has a smaller time frame, the weighing lead to a significant decline in performance. The stability experiment showed that the backbone progeny size might be more sensitive to changes in the underlying data compared to Indegree and PageRank; however, with a cutoff year all metrics were relatively stable with only minor deviations from the mean.

The manual annotation showed that there was some discrepancy between the definition of seminality used in this thesis and the gold standards used, with many additional publications being deemed seminal.

However, the overall trend is that random retrieval is outperformed, indicating that there is some correlation between being a fellow and the metrics used.

The metrics were not strongly correlated, and the logistic regression showed that we might be able to improve the overall performance by combining several metrics.

Overall, the results indicate that there is latent information in citation networks regarding the seminality of publications and methods that go beyond counting citations could possibly be useful additions to existing metrics.

5.2 Discussion

All metrics showed improvement in the APS dataset when filtering out older publications. A likely reason for this is that the time frame of the fellow program (1980–) is much smaller than the entire dataset's (1893–). This might also explain why the age filtering had a much smaller effect on the metrics in the AAN dataset, where most publications have been published in the last two decades.

The backbone graph exhibited several structural problems. All publications, except the oldest ones, have exactly one outgoing edge, resulting in a set of unary trees. Since these are time-ordered, with each edge pointing backwards in time, the potential progeny size increases with the age of a publication. This is counteracted in the weighted version of the backbone progeny size, with effects similar to the age cutoff. A fundamental problem with the fact that all publications have an outgoing edge is that even if a truly seminal publication with a large progeny size is found, it will still have an outgoing edge to another publication with an even bigger progeny size. In order to prevent this we would need to improve the selection of backbone nodes by e.g. introducing a threshold for the similarity measures, making sure all edges are meaningful.[1] In the case of J.C. Slater in Table 4.7 the edges formed a chain between his own publications, with the last one being a sink with no outgoing edge. This similarity could be due to the fact that the publications have many citations in common or that they are frequently co-cited, which can easily happen when an author publishes many papers on the same, or similar, topics. If we wanted to get rid of this phenomenon altogether we could filter out self-citations on an author level; completely diregarding any references an author has to his/her own publications.

The lower stability shown by the backbone progeny size might be due to the fact that a single backbone edge can significantly boost a node's overall progeny size. We also noted that there was high variance for short

[1] The backbone node selection process is defined in equation 2.6.

lists, which is not very surprising. For example, for lists of length 10, each fellow publication gives us an extra 0.1 (10%) precision. Since over 30% of all publications are written by fellows in the APS data some spikes are to be expected.

In the AAN data a few publications had a much higher betweenness centrality than all other publications and these were often fellow publications. A possible explanation for this is that AAN consists of a more distinct subject area where publications with high betweenness centrality more easily stand out. In a large dataset with multiple subdisciplines, seminal interdisciplinary publications might be overshadowed by other publications since a high betweenness centrality is easier to obtain. A single link from an otherwise disjoint cluster could boost a publication's betweenness centrality significantly. The general problem is that subject areas can differ a lot in size and citation culture, making it difficult to put the centrality in perspective.

Chen et al. [34] also observed a high correlation between $\sqrt{betweenness \times burstiness}$ and $\sqrt{betweenness \times burstiness \times indegree}$, indicating that $\sqrt{betweenness \times burstiness}$ should be used, for simplicity. However, Chen et al. [34] observed a much lower correlation between Burstiness and Betweenness, stating that they are almost independent measures despite their common connection to citation count. Our tests, which used larger datasets, showed a much higher, but still relatively low, correlation.

The DCG scores followed the same trend as their corresponding Precision scores with no apparent spikes, indicating that there were no significant differences in the ordering of the found fellow articles. This is most likely due to the fact that the DCG calculations are not fully utilized with binary relevance scores.

The logistic regression was very sensitive to changes in the variables. If PageRank was excluded from the regression the top performing Logit in the APS data (Figure 4.1) dropped below Random Retrieval. This is most likely due to the underlying estimation method, MLE, which maximizes the likelihood for all data rather than for the subsets evaluated by *Precision @ n*. The logistic regression's goodness of fit was low for both datasets. Looking at the class probabilites we see that they quickly become very low, further indicating that the regression failed to fit the data well. Some experiments were performed where combinations of variables were added, but the regression often failed to converge, most likely due to multicollinearity. Overall, the logistic regression showed promise, but further experiments are needed; new independent variables can be introduced and different estimation methods can be tested.

5.2.1 Time complexities

One major advantage of graph-related metrics in contrast to e.g. full text analysis is that it is fast. All of the metrics that have been used in this thesis are only dependent on the number of publications (nodes) and the number of references between publications (edges) allowing us to quickly calculate the metrics for millions of nodes. Calculating the backbone can be time-consuming, but since only local calculations are made, where a node along with its predecessors and successors is all that is required to calculate its backbone node, it can be parallelized. A summary of the time complexities for the network algorithms is shown in Table 5.1.

Table 5.1: Time complexities of the network algorithms. $|V|$ is the number of vertices, $|E|$ is the number of edges.

Metric	**Time complexity**	**Notes**
Degree centrality	$O(\|E\|)$	
Betweenness centrality	$O(\|V\|\|E\|)$ [59]	For unweighted graphs.
Backbone calculation	$O(k\|V\|)$	$k = (\max(p, c))^3$ where p and c are the maximum outdegree and indegree respectively. We compute the similarity for all pairs of a node's predecessors. The predecessor calculations are the random walks described in section 2.2.1.
Progeny size calculation	$O(\|V\|(\|V\| + \|E\|))$	Assumes a graph traversal algorithm from each node, which can be improved for DAGs where traversals can be combined by traversing in a topological ordering of the nodes.
PageRank	$O(k(\|V\| + \|E\|))$	k is the number of iterations and depends on the desired precision.

In theory, the upper bound of $|E|$ is $|V|^2$ but for citation networks the value of $|E|$ is the number of publications times the average number of references which means that, assuming the average number of references can be considered a constant, $|E|$ is $O(|V|)$. For co-citation graphs, which are used for the betweenness calculations, we get a much higher $|E|$. Each reference list becomes a complete subgraph where all of its nodes have an edge to all other node. For example, the APS citation graph had about 6 million edges and the corresponding co-citation graph had about 33 million edges.

5.2.2 Sources of error

5.2.2.1 Ranking

The drawback of using a ranking based evaluation is that information regarding the magnitude of the values are lost. For example, even if there is only one well-cited publication, all metrics will be given a ranking. The assumption being that our datasets contain enough data to create a meaningful ranking for most of the publications. *Precision @ n* partly protects us against evaluating scores that show no meaningful difference (e.g a citation count of 1 versus a citation count of 2) by only including a fraction of the results.

5.2.2.2 Fellows

Using fellows as a gold standard is based on the assumptions that all fellows are elected because of their contributions and that all of their articles are part of this contribution. In reality, some fellows might have been chosen for reasons beyond scientific excellence and only a few of their publications might relate to the nomination. Since only members can be elected, a large number of Type II errors (false negatives) where seminal publications are annotated as non-seminal will be made. An alternative approach would be to use a multitude of criteria, such as awards, nominations, or popularity and let domain experts manually classify publications as either seminal or non-seminal. This could produce a very useful gold standard since one could argue that whatever domain experts say is seminal, is in fact seminal.

5.2.2.3 Wikipedia

Wikipedia was used in order to formulate our own opionion on whether publications were seminal or not in Tables 4.7 and 4.8. The question is, can Wikipedia be used as a proxy for domain expertise? In some instances we can make objective observations, such as checking which prizes researchers have won or been nominated for, but stating that someone has made major contributions to a field is subjective. However, due to the popularity of Wikipedia, one could argue that the opinions expressed there reflect the opinions of the majority.

Another issue is that the incompleteness of Wikipedia will lead to false negatives. This is especially true for the AAN dataset since the field of computational linguistics is much smaller than the physics fields represented in the APS dataset.

One benefit of this approach is that it provides us with a way to do a sanity check, making sure our results are not completely random.

5.2.2.4 Disambiguation

When building the gold standard sets using the lists of fellows we made the assumption that authors in the datasets which had the same name as a fellow were the same person. This assumption is a possible source of error that might lead to authors being incorrectly annotated as fellows. However, name disambiguation is a difficult task and requires a lot of considerations that are beyond the scope of this thesis. As previously mentioned we often had access to the initials of middle names, which reduces the risk of name collisions.

Another problem with this approach is that authors that change names or use variations of their name when publishing articles might be incorrectly assumed to be different people.

5.2.3 Ethics

There are two main ethical aspects to consider. First, we must respect the usage rights of the data we are using. Especially in recent years companies have restricted access to their data and limited what we are allowed to do with it. For example, many sites do not allow you to crawl their data.

Second, with the ”rich get richer” phenomenon previously discussed, a metric can easily become a self-fulfilling prophecy since increased visibility more or less automatically leads to more coverage and more citations. This means that large companies with large searchable databases have a responsibility to stay unbiased, not directly affecting the visibility of research. This is a difficult task, but a good start is to minimize the effects of cheating behavior. For example, there are people that buy citations or artificially increase their citation count by cross-citing their own work with no purpose beyond increasing the count. The backbone graph's $s_{ii'}^{\text{auth}}$ component, which represents similarity as seen from the author's perspective, is subject to abuse. Since it is based on references, a person with intimate knowledge of the citation graph could attempt to boost their similarity score with selected publications. The effects of this can be minimized by e.g. setting f, the relative weight between $s_{ii'}^{\text{auth}}$ and $s_{ii'}^{\text{read}}$, to a value above 0.5, thus lowering the impact of $s_{ii'}^{\text{auth}}$.

5.2.4 Future work

In order to build a complete system for finding seminal publications several components not included here should be implemented. This includes disambiguation, different levels of pruning in the preprocessing, community detection, and full text analysis. Full text analysis could lead to a more nuanced view of the citations themselves, e.g. via sentiment analysis which could identify strictly negative citations, or by counting the number of times each reference is cited.

The backbone algorithm could be customized for different needs, such as discarding self-citations on an author level in order to focus on how authors spread ideas. The weighted variant would need to be studied more before use, finding appropriate weighing schemes for different graphs and goals. The same is true for the parameters of the burstiness algorithm which should be adapted to fit whichever view of burstiness is needed, and the data itself. For example, we might want to only consider very intense bursts, which is accomplished by increasing the γ parameter.

In order to fully utilize the power of betweenness centralities we need to introduce more preprocessing in order to overcome overshadowing problems. Dividing the data into distinct subject areas using e.g. metadata or community detection could be done as an intermediary step in order to measure a node's betweenness on a more local scale by calculating its centrality relative to its own community. One can filter out an entire graph on which to make calculations or simply identify a subdiscipline in order to normalize all node centralities within their respective areas.

The parameters of the Burstiness algorithm needs to be considered before use, especially when multiplied with betweenness centrality. If the threshold for burstiness is too low there might be too much noise, and if it is too high there may be too many publications that get a burstiness value of 0, canceling out the betweenness values in the geometric mean. If low thresholds are used it might be useful to include additional states, allowing us to differentiate between different levels of burstiness.

A goal for the entire research area of citation analysis should be to collect large amounts of quality data, along with proper gold standards, and make them free to use for research purposes. In an ideal scenario there would be free access to large and well maintained datasets that researchers could reuse, making it easy to compare results. There are already many well-used datasets out there, but they often contain inconsistencies, require a lot of preprocessing, and have very little metadata available.

Appendix A

Decision matrix calculations

Each dataset was scored on a set of attributes on a scale from 0-2 (Weak, Neutral, Good). The final score of a dataset is calculated as $\sum_a s_a \cdot w_a$ where a is an attribute, s_a is the attribute score for the dataset (0-2) and w_a is the attribute weight. The weight calculations, the attributes used, and an approximate description of their scale are described below:

A: Size (paper and reference counts)

- **Weak** = Very small or very sparse network.
- **Neutral** = Moderately large network. Average outdegree at least > 1.
- **Good** = Large network with many edges.[1]

B: Extracted metadata

- **Weak** = No metadata or no references in metadata
- **Neutral** = Some missing metadata
- **Good** = (Almost) all useful metadata available

C: Self-contained topics (there are few references to unknown publications outside of the dataset)

- **Weak** = Almost no internal edges
- **Neutral** = A moderate average indegree. Depends on actual outdegree.
- **Good** = More or less self-contained with a large average indegree close to the actual outdegree.

[1] Note that expected average outdegree is different for different subject areas

D: Disambiguated (authors, titles, venues etc. have been merged when they are most likely the same entity)

- **Weak** = No disambiguation and many duplicates
- **Neutral** = No disambiguation and few duplicates
- **Good** = Some disambiguation and few duplicates

E: Time frame

- **Weak** = Very short time frame or temporal inconsistencies (e.g. bad coverage of certain time periods)
- **Neutral** = Moderate time frame. Enough to make temporal analysis. C.a. 5-10 years, depending on the subject area.
- **Good** = Large time frame. Decades.

The decision matrix is calculated by choosing between each pair of attributes. The value of M_{ij} is either i or j, whichever is considered to be more important. The weight of each attribute is then the amount of times it was chosen. The weights are normalized by setting their sum to be 100. The decision matrix for the above attributes is:

$$\begin{pmatrix}
\begin{array}{cccccc}
 & \mathbf{A} & \mathbf{B} & \mathbf{C} & \mathbf{D} & \mathbf{E} \\
\mathbf{A} & & \multicolumn{1}{|c}{B} & C & A & A \\ \cline{3-3}
\mathbf{B} & & & \multicolumn{1}{|c}{C} & B & B \\ \cline{4-4}
\mathbf{C} & & & & \multicolumn{1}{|c}{C} & C \\ \cline{5-5}
\mathbf{D} & & & & & \multicolumn{1}{|c}{D} \\ \cline{6-6}
\mathbf{E} & & & & &
\end{array}
\end{pmatrix}$$

$$w_A = 20, w_B = 30, w_C = 40, w_D = 10, w_E = 0$$

Attribute E (time frame) received a weight of 0, meaning that the time frame had no effect on the scores. Hence the time frame is considered to be the least important attribute although any discussion regarding the usability of a dataset that uses the decision matrix score as an indicator should also discuss the time frame of the dataset.

Calculations for each individual dataset

$$\begin{aligned}
Score &= s_A \cdot w_A + s_B \cdot w_B + s_C \cdot w_C + s_D \cdot w_D + s_E \cdot w_E \\
AAN : 160 &= \quad 1 \cdot 20 \;+\; 2 \cdot 30 \;+\; 2 \cdot 40 \;+\; 0 \;+\; 0 \\
APS : 140 &= \quad 2 \cdot 20 \;+\; 2 \cdot 30 \;+\; 1 \cdot 40 \;+\; 0 \;+\; 0 \\
KDD : 130 &= \quad 1 \cdot 20 \;+\; 1 \cdot 30 \;+\; 2 \cdot 40 \;+\; 0 \;+\; 0 \\
Scopus : 130 &= \quad 1 \cdot 20 \;+\; 2 \cdot 30 \;+\; 1 \cdot 40 \;+\; 1 \cdot 10 \;+\; 0 \\
arXiv : 80 &= \quad 2 \cdot 20 \;+\; 0 \;+\; 1 \cdot 40 \;+\; 0 \;+\; 0 \\
CORE : 40 &= \quad 2 \cdot 20 \;+\; 0 \;+\; 0 \;+\; 0 \;+\; 0
\end{aligned}$$

Bibliography

[1] A. E. Jinha. “Article 50 million: an estimate of the number of scholarly articles in existence”. In: *Learned Publishing* 23.3 (2010), pp. 258–263.

[2] J. E. Hirsch. “An index to quantify an individual’s scientific research output”. In: *Proceedings of the National academy of Sciences of the United States of America* 102.46 (2005), pp. 16569–16572.

[3] E. Garfield et al. “Citation analysis as a tool in journal evaluation”. In: American Association for the Advancement of Science. 1972.

[4] Derek de Solla Price. “A general theory of bibliometric and other cumulative advantage processes”. In: *Journal of the American society for Information science* 27.5 (1976), pp. 292–306.

[5] Yuan An, Jeannette Janssen, and Evangelos E Milios. “Characterizing and mining the citation graph of the computer science literature”. In: *Knowledge and Information Systems* 6.6 (2004), pp. 664–678.

[6] Paul L Krapivsky, Sidney Redner, and Francois Leyvraz. “Connectivity of growing random networks”. In: *Physical review letters* 85.21 (2000), p. 4629.

[7] M. E. J. Newman. “The first-mover advantage in scientific publication”. In: *EPL (Europhysics Letters)* 86.6 (2009), p. 68001.

[8] X. Zhu et al. “Measuring academic influence: Not all citations are equal”. In: *Journal of the Association for Information Science and Technology* 66.2 (Feb. 2015), pp. 408–427.

[9] Simone Teufel, Advaith Siddharthan, and Dan Tidhar. “Automatic classification of citation function”. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2006, pp. 103–110.

[10] Elsevier B.V. *Scopus*. `http://www.scopus.com/`. [Online; Accessed: 2015-03-10; Login required]. 2015.

[11] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008.

[12] Peixiang Zhao, Jiawei Han, and Yizhou Sun. “P-Rank: a comprehensive structural similarity measure over information networks”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 553–562.

[13] Hongbo Deng et al. “Probabilistic topic models with biased propagation on heterogeneous information networks”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2011, pp. 1271–1279.

[14] Hongbo Deng et al. “Modeling and exploiting heterogeneous bibliographic networks for expertise ranking”. In: *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*. ACM. 2012, pp. 71–80.

[15] Ying Ding, Ronald Rousseau, and Dietmar Wolfram. *Measuring Scholarly Impact*. Springer, 2014.

[16] Linton C Freeman. “A set of measures of centrality based on betweenness”. In: *Sociometry* (1977), pp. 35–41.

[17] Loet Leydesdorff. “Betweenness centrality as an indicator of the interdisciplinarity of scientific journals”. In: *Journal of the American Society for Information Science and Technology* 58.9 (2007), pp. 1303–1319.

[18] Jae Dong Noh and Heiko Rieger. “Random walks on complex networks”. In: *Physical review letters* 92.11 (2004), p. 118701.

[19] Tien-Dzung Tran and Yung-Keun Kwon. “Hierarchical closeness efficiently predicts disease genes in a directed signaling network”. In: *Computational biology and chemistry* 53 (2014), pp. 191–197.

[20] Naoki Shibata, Yuya Kajikawa, and Katsumori Matsushima. “Topological analysis of citation networks to discover the future core articles”. In: *Journal of the American Society for Information Science and Technology* 58.6 (2007), pp. 872–882.

[21] Phillip Bonacich. “Power and centrality: A family of measures”. In: *American journal of sociology* (1987), pp. 1170–1182.

[22] Jon M Kleinberg. “Authoritative sources in a hyperlinked environment”. In: *Journal of the ACM (JACM)* 46.5 (1999), pp. 604–632.

[23] Lawrence Page et al. *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report 1999-66. Stanford InfoLab, Nov. 1999. URL: `http://ilpubs.stanford.edu:8090/422/`.

[24] Carl T Bergstrom, Jevin D West, and Marc A Wiseman. “The Eigenfactor™ metrics”. In: *The Journal of Neuroscience* 28.45 (2008), pp. 11433–11434.

[25] Dylan Walker et al. “Ranking scientific publications using a model of network traffic”. In: *Journal of Statistical Mechanics: Theory and Experiment* 2007.06 (2007), P06010.

[26] Borja González-Pereira, Vicente P Guerrero-Bote, and Félix Moya-Anegón. “A new approach to the metric of journals’ scientific prestige: The SJR indicator”. In: *Journal of informetrics* 4.3 (2010), pp. 379–391.

[27] Peng Chen et al. “Finding scientific gems with Google’s PageRank algorithm”. In: *Journal of Informetrics* 1.1 (2007), pp. 8–15.

[28] Elizabeth A Leicht et al. “Large-scale structure of time evolving citation networks”. In: *The European Physical Journal B-Condensed Matter and Complex Systems* 59.1 (2007), pp. 75–83.

[29] Arthur P Dempster, Nan M Laird, and Donald B Rubin. “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.

[30] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.

[31] Duck-Ho Bae et al. “On Constructing Seminal Paper Genealogy”. In: *Cybernetics, IEEE Transactions on* 44.1 (2014), pp. 54–65.

[32] Stanislao Gualdi, Matus Medo, and Y-C Zhang. “Influence, originality and similarity in directed acyclic graphs”. In: *EPL (Europhysics Letters)* 96.1 (2011), p. 18004.

[33] Stanislao Gualdi, Chi Ho Yeung, and Y-C Zhang. “Tracing the evolution of physics on the backbone of citation networks”. In: *Physical Review E* 84.4 (2011), p. 046104.

[34] Chaomei Chen et al. “Towards an explanatory and computational theory of scientific discovery”. In: *Journal of Informetrics* 3.3 (2009), pp. 191–209.

[35] Jon Kleinberg. “Bursty and hierarchical structure in streams”. In: *Data Mining and Knowledge Discovery* 7.4 (2003), pp. 373–397.

[36] Hong Yu et al. “An Improved Random Walk Algorithm Based on Correlation Coefficient to Find Scientific Communities”. In: *Computer Science and Software Engineering, 2008 International Conference on*. Vol. 1. IEEE. 2008, pp. 690–693.

[37] Kalervo Järvelin and Jaana Kekäläinen. “IR evaluation methods for retrieving highly relevant documents”. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.* ACM. 2000, pp. 41–48.

[38] Chris Burges et al. “Learning to rank using gradient descent”. In: *Proceedings of the 22nd international conference on Machine learning.* ACM. 2005, pp. 89–96.

[39] Karl Pearson. “Mathematical Contributions to the Theory of Evolution.– On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs”. In: *Proceedings of the royal society of london* 60.359-367 (1896), pp. 489–498.

[40] Jevin West, Theodore Bergstrom, and Carl T Bergstrom. “Big Macs and Eigenfactor scores: Don’t let correlation coefficients fool you”. In: *Journal of the American Society for Information Science and Technology* 61.9 (2010), pp. 1800–1807.

[41] Jerome L Myers, Arnold Well, and Robert Frederick Lorch. *Research design and statistical analysis.* Routledge, 2010.

[42] Cosma Rohilla Shalizi. “Advanced data analysis from an elementary point of view”. In: *Preprint of book found at http://www.stat.cmu.edu/ cshalizi/ADAfaEPoV* (2013).

[43] DragomirR. Radev et al. “The ACL anthology network corpus”. English. In: *Language Resources and Evaluation* (2013). [Online; Accessed: 2015-04-21], pp. 1–26. ISSN: 1574-020X. DOI: 10.1007/s10579-012-9211-2. URL: http://dx.doi.org/10.1007/s10579-012-9211-2.

[44] APS Journals, American Physical Society. http://journals.aps.org/datasets. [Online; Accessed: 2015-03-10]. 2013.

[45] KDD Cup 2003, Cornell University. http://www.cs.cornell.edu/projects/kddcup/datasets.html. [Online; Accessed: 2015-03-10]. 2003.

[46] *arXiv.org.* http://arxiv.org/. [Online; Accessed: 2015-02-15].

[47] CORE, Knowledge Media Institute. http://core.ac.uk/intro/data_dumps. [Online; Accessed: 2015-03-10]. 2014.

[48] J. E. Hirsch. “Does the h index have predictive power?” In: *Proceedings of the National Academy of Sciences* 104.49 (2007), pp. 19193–19198.

[49] Ding Zhou. *Mining social documents and networks.* ProQuest, 2008.

[50] *APS Fellows.* http://www.aps.org/programs/honors/fellowships/index.cfm. Online; Accessed: 2015-04-21.

[51] *ACL Fellows.* http://aclweb.org/aclwiki/index.php?title=ACL_Fellows. Online; Accessed: 2015-04-21.

[52] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring network structure, dynamics, and function using NetworkX". In: *Proceedings of the 7th Python in Science Conference (SciPy2008).* Pasadena, CA USA, Aug. 2008, pp. 11–15.

[53] *graph-tool.* http://graph-tool.skewed.de/. Online; Accessed: 2015-04-21.

[54] *pandas.* http://pandas.pydata.org/. Online; Accessed: 2015-04-21.

[55] *Statsmodels.* http://statsmodels.sourceforge.net/. Online; Accessed: 2015-04-21.

[56] *Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies.* https://sci2.cns.iu.edu. Online; Accessed: 2015-04-21.

[57] *SequenceMatcher.* https://docs.python.org/2/library/difflib.html#difflib.SequenceMatcher. Online; Accessed: 2015-05-22.

[58] *Wikipedia.* https://www.wikipedia.org/. Online; Accessed: 2015-04-26.

[59] Ulrik Brandes. "A faster algorithm for betweenness centrality*". In: *Journal of Mathematical Sociology* 25.2 (2001), pp. 163–177.