PROBABILISTIC ARGUMENTATION SYSTEMS ENTITY-TRANSITIVE RELATION-IMPLICATION MODEL AND DOCUMENT RANKING AS AN EFFICIENT APPLICATION

by

Burak Çetin

B.Sc. in Electrical Engineering, Boğaziçi University, 2000

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Computer Engineering
Boğaziçi University

2005

PROBABILISTIC ARGUMENTATION SYSTEMS
ENTITY-TRANSITIVE RELATION-IMPLICATION MODEL
AND DOCUMENT RANKING AS AN EFFICIENT APPLICATION

APPROVED BY:

Dr. Haluk Bingöl . . . . . .
(Thesis Supervisor)

Assoc. Prof. Yağmur Denizhan . . . . . .

Assistant Prof. Tunga Güngör . . . . . .

DATE OF APPROVAL:

ABSTRACT

PROBABILISTIC ARGUMENTATION SYSTEMS
ENTITY-TRANSITIVE RELATION-IMPLICATION MODEL
AND DOCUMENT RANKING AS AN EFFICIENT APPLICATION

This work is an endeavor towards analyzing complex networks. Mainly, a link analysis ranking (LAR) algorithm will be introduced, and related background will be developed.

Firstly, we introduce a graph based model we name Entity-Transitive Relation-Implication Model (ETRI) for analyzing complex networks. The underlying mathematical model is built on Probabilistic Argumentation Systems (PAS), which are a combination of the use of propositional logic and probability theory. The ETRI model is a generic framework, capable of dealing with entities (e.g. web pages) in a network linked by a transitive relation (e.g. hypertext links). We apply ETRI modeling to the LAR problem. This is desirable because it builds on established evidential reasoning techniques using clear semantics, however a direct application involves an NP-hard problem. Thus we present a family of novel algorithms we call ETRI Support Propagation for approximations. We examine a member of these and show that it produces approximate results in finite iterations. Its iterations are linear in the number of edges of the network like PageRank. We run our algorithms on a snapshot of the CiteSeer citation network. We present a comparative study of different ranking schemes. Our studies reveal the transition of dominance from local to global influences as an important characteristic of LAR algorithms. Our algorithms give results which can be highly correlated with citation count or PageRank when parameterized correspondingly.

ÖZET

OLASILIKSAL MUHAKEME (ARGÜMANLAMA) SİSTEMLERİ
NESNE-GEÇİŞSEL İLİŞKİ-GEREKTİRME MODELİ
VE VERİMLİ UYGULAMALARI

Bu çalışma karmaşık ağların incelenmesine yönelik yapılmış bir çabadır. Temel olarak, bir bağ analizi tabanlı seviyelendirme (**BTS**) algoritması tanıtılacak ve ilgili altyapı geliştirilecektir.

Öncelikle karmaşık ağların incelenmesi için grafik tabanlı *Nesne-Geçişsel İlişki-Gerektirme* (**NGİG**) modeli tanıtılacak ve kullanımı incelenecektir. Altyapıyı oluşturan matematik model *Olasılıksal Muhakeme Sistemleri* (**OMS**) üzerinde yapılmış olup bu sistemler de matematik lojik ve olasılık teorisi üzerine kurulmuşlardır. NGİG modeli genel bir çerçeve olup bir ağ yapısı içindeki nesnelerle (örn. ağ sayfaları, makaleler) bunları bağlayan geçişsel bir bağı (örn: ağ bağları ("link"ler)) incelemek için yapılmıştır. NGİG modellemesini BTS problemi için uygulamaktayız. Bu işlem için yerleşmiş *kanıtsal sebep üretme* tekniklerini açık bir şekilde kullanmaktayız, ancak direk hesaplamalar NP-zor bir problem içermektedir. Bu sebeple yaklaşık sonuç üreten *NGİG Destek Yayılması* olarak adlandırdığımız algoritma ailesini sunmaktayız. Bunlardan bir tanesini detaylı inceleyerek, sonlu sayıda iterasyon ile yaklaşık sonuçlar ürettiğini gösteriyoruz. Her iterasyon için yapılan işlemler ağ içindeki bağ sayısı ile lineer şekilde bağlantılıdır. Algoritmalarımızı CiteSeer bilimsel atıf ağına uyguladık. Bu ağ üzerinde seviyelendirme yapılarının karşılaştırmalı sonuçlarını sunmaktayız. Çalışmamız baskınlığın küresel etkilerden yerel etkilere geçişinin temel bir BTS algoritması karakteristiği olduğunu ortaya çıkardı. Algoritmamız farklı parametreler ile kullanıldığında PageRank veya atıf sayımı ile yüksek korelasyonlu olabilen sonuçlar üretmektedir.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

$\perp$	Contradiction (inconsistency)
$\models$	Entails
$\not\models$	Does not entail
ξ	Knowledge-base in a PAS instance
$\hat{\vee}$	Noisy-or operator
Π	Set of assumption probabilities in a PAS instance
A	Set of assumptions in a PAS instance
a_i	Node assumption for vertex i
AS_P	A propositional argumentation system instance
d_0	Constant valued damping function
$dqs(h, \xi)$	Degree of quasi-support for h
$dsp(h, \xi)$	Degree of support for h
dsp_i	Degree of support of vertex i
$\hat{dsp}_i$	An approximation for dsp_i
E	Set of arcs in an ETRI
h	Hypothesis
l_{ij}	Link assumptions from vertex i to vertex j
n	Number of vertices in a network
P	Set of propositions in a PAS instance
P_i	Parents of vertex i
PAS_P	A probabilistic argumentation system instance
$pla(h, \xi)$	Plausibility of h
$QS(h, \xi)$	Quasi-support for h
$QS_A(h, \xi)$	Quasi-support for h (set of quasi-supporting scenarios)
R	Relation in an ETRI
$RSV(d^i)$	Retrieval status value for document d on i^{th} iteration
$SP(h, \xi)$	Support for h
$SP_A(h, \xi)$	Support for h (set of supporting scenarios)

V	Set of vertices in an ETRI
v_i	Proposition for vertex i
APD	Average position distance
BDD	Binary Decision Diagram
BP	Belief Propagation
DIM	Document/Information value Model
DNF	Disjunctive normal form
DRM	Document/Relevance Model
DST	Dempster Shafer Theory of Evidence
ERank-n	An n^{th} order ESP based ranking algorithm
ESP	ETRI Support Propagation
ESP-n	n^{th} order ETRI Support Propagation
ETRI	Entity-Transitive Relation-Implication
IR	Information Retrieval
LAR	Link Analysis Ranking algorithms
ME	Minimal Evidence
PAS	Probabilistic Argumentation Systems
SAT	Satisfiability problem

1. INTRODUCTION

1.1. Motivation and Essentials

Recent years have seen an increasing interest in two important fields of research. These are document ranking algorithms and complex networks. In 1998 Page and Brin (Page *et al.*, 1998) along with Kleinberg (Kleinberg, 1999) have disclosed two similar algorithms which since then have deeply affected the web experience by the creation of very successful search engines. A foremost one is the Google search engine which is run by the very authors who have discovered the algorithms (Page *et al.*, 1998).

On the other hand, the research in complex networks from authors such as Watts and Barabasi have resulted in interesting discoveries in the structure of the Web and some other well known networks, and perhaps more importantly common properties were found to be shared across complex network structures of vastly different kinds ranging from social networks to computer networks. This created an exciting prospect in understanding the essence of complexity and complex behavior.

Our work sits between the two as it resulted from an interest in analyzing complex networks and as its initial focus has dealt with link analysis ranking. Towards this end the third important component has come in the picture as our tool of analysis; Probabilistic Argumentation Systems (PAS) (Haenni *et al.*, 2000). Probabilistic reasoning is an ever getting stronger discipline. One can see today's probabilistic reasoning systems as the accumulation of decades of work, as PAS builds on Dempster Shafer Theory of Evidence (DST) (Shafer, 1976), (Shafer, 1990), (Dempster, 1968).

In our work we mainly present and analyze a link analysis ranking algorithm we name ERank-0, which extends citation count using probabilistic argumentation. The algorithm is iterative with linear time-complexity in the number of edges for each iteration (like PageRank (Page *et al.*, 1998)). This linear time-complexity is central for our work because we intend its application on very large networks, like the Web.

In coming chapters we will introduce the algorithm formally and analyze its theoretical aspects. We have used the CiteSeer (CIT) citation network for experimenting with different algorithms which has around 300 000 nodes and 1 250 000 edges. In our results we will present a comparative study of ERank-0, PageRank and citation count.

Forming the background for ERank-0 is a framework we call Entity-Transitive Relation-Implication (ETRI). ETRI is a combination of a graph model and a corresponding Probabilistic Argumentation Systems (PAS) instance (Haenni *et al.*, 2000). The idea originates in (Picard, 1998), and we essentially generalize and try to formalize that model here. ETRI is a generic model which we perceive as a tool for analyzing complex networks.

We built an efficient family of approximation algorithms on ETRI we call ETRI Support Propagation (ESP) algorithms. ESP-0 is the 0^{th} order algorithm in that family, has linear time-complexity in the number of edges, and is essentially the simplest to implement and analyze. ERank-0 is a straightforward application of ESP-0 to the ranking problem. Our theoretical analysis of ERank-0 is in ESP-0 the context.

The generic definition of the ETRI framework structure has offered us the possibility to project our results on networks of a great variety of choice as there are many such structures involving a transitive relation, links and nodes. To name a few; internet, citation networks, disease spreading/epidemics analysis, forest fire prevention optimization, software function call graphs. We hope the ESP algorithms and the ETRI framework to prove to be useful tools in the future not only for ranking for but also for different purposes such as community detection or network characterization.

1.2. Current Techniques

Two most important document ranking algorithms are citation count (in-degree) and PageRank (Page *et al.*, 1998). We focus on these, and compare them with our own algorithms within our work.

Citation count is a very old, simple, and surprisingly effective ranking algorithm. PageRank in its origin had the very same purpose like our work, “to extend citation count” to incorporate the value of the citing party.

PageRank for this purpose introduces a novel concept “random walk” and uses this as a popularity measure. We choose a different path, and perceive the situation from a DST based evidence perspective.

PageRank has seen many extension and variations since its introduction. Also there are various other link analysis ranking algorithms (Borodin *et al.*, 2005). However, PageRank and citation count still remain central and most researched algorithms, and have served our purposes well for comparing algorithms.

PAS formalism combines propositional logic with probability in the form of evidence as introduced in DST. There exists similar formalisms using first-order logic with probability theory (Laskey, 2005), (Poole, 2003). Also of relevance are credal networks (quasi-bayesian networks) which incorporate uncertainty in probabilities (Cozman, 2000). We have found PAS to be very valuable in providing a compact and clear framework which has a natural adeptness to our analysis purposes. In our opinion the concept of “evidence” as opposed to uncertain probabilities serves the problem definition more naturally.

1.3. Main Contributions

The ETRI idea is developed and introduced in (Picard, 1998), as an application model for Information Retrieval. We believe that identification of ETRI as a general network analysis tool for employing “evidential perspective” on analysis may provide beneficial for further research. To help facilitate this we develop a formal treatment of the model and try to enumerate useful applications.

Possibly our most important contribution is the introduction of ETRI Support Propagation (ESP) family of algorithms. ESP algorithms provide a rapid way to calculate close approximations in an ETRI network, a problem which is NP-hard otherwise. We provide a theoretical treatment for the first member of these algorithms (ESP-0), and cover topics such as convergence and accuracy.

ERank-0 which we base on ESP-0, appeared to be a promising algorithm in our experiments on CiteSeer citation network. ERank-0 is desirable because it builds with clear semantics on evidential reasoning in the sense of DST without making any unjustified assumptions, and it is also capable of scaling to very large collections.

Our study reveals well a characteristic property of different ranking algorithms. It is the transition of dominance from local to global influences on ranks. We believe our work is valuable in that, we were able to track and exemplify these changes using different parameters for on algorithm runs and making comparative studies of these different settings.

1.4. Experimental Setup

For evaluating our work we have mainly worked on artificially generated scale free networks and later the CiteSeer (CIT) citation network which had around 300 000 nodes and 1 250 000 edges.

We have used the open-source JUNG codebase (JUN) for manipulating network data structures, and eventually rewrote some core code to fit our performance and memory limitations.

We set up an SQL database server, and stored content info on tables. This enabled full-text queries on the whole network, eventually creating an experimental search engine.

Additionally, the background foundation required coding of an ETRI based PAS implementation. The work-horse for the implementation has been the open-source JavaBDD codebase (JBD), and indirectly the CUDD BDD implementation (CUD).

1.5. Organization of the Document

The rest of the document is organized as follows. Chapter 2 contains all the related literature survey and preliminary offered. Three topics are treated, these are PAS, complex networks, and link analysis ranking algorithms. Section 2.4 contains important additional background that is used through-out the text.

Chapter 3 formally introduces the ETRI model, is a short treatment on its uses. Chapter 4 explores the use of ETRI as LAR algorithm, and introduces the Minimal Evidence (ME) concept, along with ArgRank scheme.

Chapter 5 is devoted to efficient approximations on the ETRI network. The ESP family of algorithm are presented here, and ESP-0 is examined closely. ERank-0 is defined in this chapter.

Chapter 6, contains all the experimental results we have obtained. These include, general analyses of the CiteSeer citation network including the rank distributions on nodes. The accuracy of ERank-0 as an approximation to ArgRank is examined. Comparative plots are presented to investigate similarities in ranking schemes. A new measure called “average position distance” is introduced and employed. Top ranking documents are listed, and found to be different between different ranking schemes. The chapter concludes with exemplary query results.

The conclusion and future work sections are in chapter 7. Appendix A contains the proofs for all the theorems in this work. Appendix B is a brief treatment of the PAS-ETRI implementation created as background work.

2. RELATED RESEARCH AND OTHER PRELIMINARIES

In this chapter we try to develop the foundation for three main topics of interest in our work, and present further the mathematical background necessary for the rest of the text.

In section 2.1, we focus on Probabilistic Argumentation Systems which form the backbone of our work. A complete formal introduction is not necessary as it is available elsewhere (Haenni *et al.*, 2000), instead we try to develop a useful intuition using extended examples and formally present only what is sufficient.

We conceive our work as an attempt to develop a useful tool for analyzing complex networks. So in section 2.2, there is a brief survey of the concept which lays the foundation for the experimental results we present in chapter 6.

Ultimately, in this text we develop an efficient link analysis ranking algorithm (LAR) based on PAS, as our prime application of the ETRI model of complex networks and the ETRI Support Propagation (ESP) algorithm. Thus, we present in section 2.3 an overview of LAR algorithms primarily focusing on PageRank, which is the closest similar algorithm in the literature.

Section 2.4 develops the mathematical background used through-out the text.

2.1. Probabilistic Argumentation Systems (PAS)

In our work, we have perceived PAS from a complex network analysis perspective, and this will be our way of introducing the theory. In this section we will introduce the necessary terminology, and the concepts on which we later on build our work.

This will allow us to depict an outline of the theory which should allow the reader to develop an understanding on the general capabilities of PAS. On the other hand, we will omit some important aspects; like non-monotonicity, and topics like efficient general purpose implementations which did not have practical importance for us. The interested reader is advised to consult the references (Haenni *et al.*, 2000) for a comprehensive treatment of PAS. The reader should note that the theory appears to have been modified from the version in (Kohlas and Haenni, 1996).

Simply put, PAS provide a framework in which, variables with uncertainty relating to a problem along with their relations with other variables are encoded in a knowledgebase. The uncertainty factor is introduced with the use of random variables. Then, solutions are found or hypotheses are justified using arguments in line with the knowledgebase, and the associated confidence in them is derived using our combined confidence in those supporting arguments.

As is often the case for PAS, we will focus on those knowledge bases encoded using propositional logic. It is easier and more compact to deal with PAS in this case (Haenni *et al.*, 2000).

We will start with introducing the propositional argumentation systems. Then, we will present an example propositional system, and informally develop underlying ideas in section 2.1.2. Then we will make a more formal introduction of the fundamental concepts. Though, this section is consistent and presents enough basic information, the interested reader is strongly advised to refer also to (Haenni *et al.*, 2000) for a wider background on the topic. Note that, on PAS related topics, we share the terminology with (Haenni *et al.*, 2000) for preserving consistency. In section 2.1.4 we introduce the probabilistic aspect of PAS, and then we attempt to develop the underlying intuition with an extended example. In sections 2.1.5 and 2.1.6 we review an application of the PAS framework on Information Retrieval (Picard, 2000), which will constitute a very important starting point for our work.

2.1.1 Representing Uncertainty Using Propositional Logic

For many, the impression is that it is not possible to express uncertainty using propositional logic. However this is no longer the case when assumptions are introduced as a new class of propositions.

In this framework, a proposition is taken as an undoubted fact. For example, "v_1" can signify "it is sunny". It is possible to construct simple certain (undoubted) rules using propositions; "$v_2 \rightarrow v_1$".

When assumptions are introduced, propositional sentences can be used to express uncertainty. For example, if "v_3" is "it will rain tomorrow", than an uncertain fact can be expressed as; "$a_1 \rightarrow v_3$" or in English "if assumption a_1 is true it will rain tomorrow". When we consider that a_1 is a random variable, what we effectively get is "it may rain tomorrow".

Note here that propositions are used in two different ways. Depending on the context, they are used both to refer to the corresponding statement "v_1", or that the statement is true ("v_1 = T"). In this sense, "$\neg v_1$" is used as "v_1 = F" as in "it is not sunny".

It is possible to create an uncertain rule. The rule "$v_1 \rightarrow v_2$" can be made an uncertain rule, when we write "$a_1 \rightarrow (v_1 \rightarrow v_2)$". Then this rule can affect inference results only if a_1 is known to be true, and a_1 being an assumption thus inserts the uncertainty in the rule.

From here onwards, we will use "v_1, v_2, v_3 ... v_n" to denote normal propositions (i.e. facts) and "a_1, a_2, a_4 ... a_n" to denote assumptions. This is slightly a different terminology compared to previous literature on the topic, yet it serves our purposes better when ETRI terminology is included.

What we have seen so far are basic constructs. However, it is possible to create rules with any desired complexity to the extent that it is possible to encode them using propositional logic. Note that, the differences between assumptions and normal

propositions are differences only in our perception, so that they do not create a difference in the way we deal with them using propositional logic.

Table 2.1. Knowledge representation in PAS

Type of knowledge	Logical representation	Natural language equivalent
a fact	v_1	"v_1 is true"
a simple rule	$v_1 \rightarrow v_2$	"v_1 entails v_2"
an uncertain fact	$a_1 \rightarrow v_1$	"if assumption a_1 is true, then v_1 is true"
a simple uncertain rule	$a_1 \rightarrow (v_1 \rightarrow v_2)$ $\leftrightarrow (a_1 \wedge v_1) \rightarrow v_2$	"if assumption a_1 is true, then v_1 entails v_2"

2.1.2 Introductory Example

We will present here a simple introductory example to facilitate the following theoretical discussion. This will also serve as an informal introduction to such concepts as; knowledge-base, hypothesis, scenarios, and others. We will introduce different ways of viewing the situation, to help develop an intuition of the underlying systematic.

The concept "knowledge-base", is frequently used in the AI literature, for example to describe an agent's perception of the outer world. Here we will employ it to refer to a conjunction of propositional sentences, which together will form the knowledge-base of a PAS instance.

Consider the knowledge-base (set representation)

$$\xi = \{\ a_1 \rightarrow v_1\ ,\ a_2 \rightarrow v_2\ ,\ v_2 \rightarrow (a_3 \rightarrow v_1),\ a_4 \rightarrow \neg\, v_1\ \} \tag{2.1}$$

which is equivalent to (sentence representation):

$$\xi = (a_1 \rightarrow v_1) \wedge (a_2 \rightarrow v_2) \wedge (v_2 \rightarrow (a_3 \rightarrow v_1)) \wedge (a_4 \rightarrow \neg\, v_1) \tag{2.2}$$

We initially can observe that, this sentence is satisfiable, and also contains a contradiction; a_1 and a_4 can not be true at the same time.

Assume that we would like to investigate the situation of the proposition "v_1" with respect to the assumptions we can make (or observe in a system for that matter). This eventually creates our hypothesis "$h = v_1$".

See Table 2.2 for the general picture of our example PAS instance containing all the 16 different assignments on our four assumptions. These are called "scenarios", and are an essential part of the PAS theory.

On this table, we can see two of the fundamental concepts relating to PAS; the sets of quasi-supporting $QS_A(h, \xi)$ and supporting scenarios $SP_A(h, \xi)$. Simply put, a scenario is said to be supporting, if the hypothesis can be shown to be true with the assignments a scenario has (a truth value for each and every assumption), and that it does not contain inconsistency. For quasi-supporting scenarios the consistency requirement is dropped.

Table 2.2. Scenarios for the example propositional argumentation system with $h=v_1$

Scenario #	a_1	a_2	a_3	a_4	$h=v_1$	$s \in QS_A(h, \xi)$	$s \in SP_A(h, \xi)$	$s \in SP_A(\perp, \xi)$
s_1	0	0	0	0	0			
s_2	0	0	0	1	0			
s_3	0	0	1	0	0			
s_4	0	0	1	1	0			
s_5	0	1	0	0	0			
s_6	0	1	0	1	0			
s_7	0	1	1	0	1	X	X	
s_8	0	1	1	1	$\perp$	X		X
s_9	1	0	0	0	1	X	X	
s_{10}	1	0	0	1	$\perp$	X		X
s_{11}	1	0	1	0	1	X	X	
s_{12}	1	0	1	1	$\perp$	X		X
s_{13}	1	1	0	0	1	X	X	
s_{14}	1	1	0	1	$\perp$	X		X
s_{15}	1	1	1	0	1	X	X	
s_{16}	1	1	1	1	$\perp$	X		X

On Table 2.2 an "X" specifies that the membership relation specified on top of the column holds for that particular row.

We see on the table that, the scenarios s_8, s_{10}, s_{12}, s_{14} and s_{16} generate contradictions but support h, so they are included in the quasi-support. On the other hand for $SP_A(h, \boldsymbol{\xi})$ we only have the consistent supporting scenarios, because intuitively they are the ones which conform to "reality" for the model and thus are worthy of consideration. The quasi-support is important only from a computational point of view.

As can be gathered from the table, additions of new variables would result in exponentially big tables, and this is why we can use term representations instead of scenarios.

As we can directly infer from the knowledge base the quasi-support is:

$$QS(h, \xi) = a_1 \vee (a_2 \wedge a_3) \tag{2.3}$$

Or we can use the set representation:

$$QS(h, \xi) = \{\ a_1, a_2 \wedge a_3\ \} \tag{2.4}$$

Here "a_1" and "$a_2 \wedge a_3$" are the quasi-supporting arguments.

We note that these can be made to include contradictions, and supporting arguments should not contain contradictions. Excluding the inconsistent scenarios we get the support for h:

$$SP(h, \xi) = (a_1 \wedge \neg a_4) \vee (a_2 \wedge a_3 \wedge \neg a_4) \tag{2.5}$$

Or:

$$SP(h, \xi) = \{a_1 \wedge \neg a_4, a_2 \wedge a_3 \wedge \neg a_4\ \} \tag{2.6}$$

Note that Eq.(2.6) shows the minimal supporting arguments. We could also write:

$$SP(h, \xi) = \{a_1 \wedge \neg a_4 \wedge a_2, a_1 \wedge \neg a_4 \wedge \neg a_2, a_2 \wedge a_3 \wedge \neg a_4\ \} \tag{2.7}$$

Eq.(2.7) is also correct as it correctly specifies support for *h*. See that it yields exactly the same scenario table as in Table 2.2. But it is not the minimal support anymore as it contains unnecessarily long terms. The same is valid between quasi-support and minimal quasi-support.

2.1.3 Fundamentals on Propositional Argumentation Systems

In this section we will present a more formal introduction to propositional argumentation systems. However, we will reveal as much as what is necessary for understanding this work, the interested reader should consult (Haenni *et al.*, 2000) for a general purpose treatment.

Definition 2.1. *Let A and P be two disjoint sets of propositions. If ξ is a propositional sentence in the propositional language created using $A \cup P$, then a triple $AS_P = (\xi, P, A)$ is called a* ***propositional argumentation system****. ξ is called the knowledge base of AS_P.*

A **literal** is a proposition. A **clause** is a disjunction of literals, whereas a **term** is a conjunction of literals.

A **hypothesis** is a sentence in the language $L_{A \cup P}$ based on propositions in A ∪ P. The conditions under which hypothesis is true or false is a central point of focus.

Definition 2.2. *Let ξ and h be two propositional sentences in $L_{A \cup P}$. Consider a term α from $L_{A \cup P}$ and that $\alpha \not\models \bot$. α is called a*

(1) ***quasi-supporting*** *argument for h relative to ξ, if $\alpha \wedge \xi \models h$*

(2) ***supporting argument*** *for h relative to ξ, if $\alpha \wedge \xi \models h$ and $\forall \alpha' \supseteq \alpha, \alpha' \wedge \xi \not\models \bot$ where α' is a term from $L_{A \cup P}$ and $\alpha' \not\models \bot$.*

Note that these definitions may appear complicated because we have by-passed the notion of scenarios in defining them. For a clearer and more intuitive introduction the reader can consult (Haenni *et al.*, 2000).

Using the definitions of arguments as above, we define the quasi-support and support relating to a hypothesis as:

Definition 2.3. *Let ξ and h be two propositional sentences in $L_{A \cup P}$, α_i a term from $L_{A \cup P}$ and that $\alpha \not\models \bot$. Then we define:*

*(1) the **quasi-support** for h relative to ξ is the disjunction of all quasi-supporting arguments:*

$$QS(h, \xi) = \bigvee\{ \alpha_i : \alpha_i \wedge \xi \models h \} \tag{2.8}$$

*(2) the **support** for h relative to ξ is the disjunction of all supporting arguments:*

$$SP(h, \xi) = \bigvee\{ \alpha_i : \alpha_i \wedge \xi \models h \text{ and } \forall \alpha' \supseteq \alpha_i, \alpha' \wedge \xi \not\models \bot \} \tag{2.9}$$

where α' is a term from $L_{A \cup P}$ and $\alpha' \not\models \bot$.

Note that, we actually do not need all the arguments to define the support and quasi-support. What is essentially needed are the minimal arguments. If for all terms in a set there is no shorter term $\alpha' \subseteq \alpha$ contained, then it is called a **minimal term representation**. The minimal term representations for support and quasi-support are equally good, and are called **minimal support** and **minimal quasi-support** respectively.

2.1.4 Extending to Probabilistic Argumentation Systems

We have dealt so far only with the qualitative aspect of the systems, now we will introduce the probabilistic part.

Definition 2.4. *Given a propositional argumentation system $AS_P = (\xi, P, A)$, and a set Π of probabilities assigned to propositions in A, then the quadruple $PAS_P = (\xi, P, A, \Pi)$ is a called a* ***probabilistic argumentation system (PAS)****.*

With the probabilistic aspect introduced, it is possible to do quantitative as well as qualitative analysis on hypotheses. Note that, all the random variables (probabilities of assumptions) are assumed to be stochastically independent.

In this framework, degree of support and degree of quasi-support are two fundamental figures. Recall the concept of a scenario from the example in 2.1.2. In that sense, the degree of quasi-support for a scenario is simply the multiplication of the probabilities of its assignments; $p(a_i)$ if an assumption a_i is 1 and $1\text{-}p(a_i)$ if it is 0. Then simply, degree of quasi-support for a hypothesis is the sum of the quasi-support of its scenarios. Note that, this can happen because the random variables are assumed to be mutually independent.

Quoting from (Picard, 2000), "The degree of support is defined as the probability of the quasi-support, conditioned on the fact that the knowledge base is satisfiable (not contradictory)."

Definition 2.5 *Let ξ and h be two propositional sentences in $L_{A \cup P}$.*

(1) the ***degree of quasi-support*** *of h relative to ξ is:*

$$dqs(h,\xi) = p\big(QS(h,\xi)\big) \tag{2.10}$$

(2) the ***degree of support*** *of h relative to ξ is:*

$$dsp(h,\xi) = p\big(SP(h,\xi) \mid \neg QS(\bot,\xi)\big) \tag{2.11}$$

For calculating *dsp* values, we see that:

$$\begin{aligned} dsp(h, \xi) &= p\big(SP(h,\xi) \mid \neg QS(\bot,\xi)\big) \\ &= p\big(\big(QS(h,\xi) \wedge \neg QS(\bot,\xi)\big) \mid \neg QS(\bot,\xi)\big) \\ &= \frac{p\big(QS(h,\xi) \wedge \neg QS(\bot,\xi)\big)}{p(\neg QS(\bot,\xi))} \\ &= \frac{p\big(QS(h,\xi)\big) - p\big(QS(\bot,\xi)\big)}{1 - p(QS(\bot,\xi))} \end{aligned} \tag{2.12}$$

The reader can consult (Haenni *et al.*, 2000) for a theoretical treatment on the calculation of dsp and dqs values, we will demonstrate it on an example in the following section.

Degree of quasi-support corresponds to unnormalized belief in Dempster-Shafer theory of evidence (DST) (Shafer, 1976). Degree of support corresponds to normalized belief.

In PAS dsp values correspond to probabilities (Haenni *et al.*, 2000). In this sense, PAS create the bridge between DST and probability theory. For example, given the "prior" probabilities of assumptions, the value $dsp(h, \xi)$ is interpreted to be the posterior probability that h is true.

There is also another value of interest in this regard. The plausibility (*pla*) of an hypothesis *h* is:

$$pla(h, \xi) = 1 - dsp(\neg h, \xi) \tag{2.13}$$

In a sense it represents an upper-bound for the probability of h, based on the current information. Note that, PAS give a non-monotonic system of evaluation, so added

information to the knowledge based can increase or decrease degrees of support without being committed in one direction.

Note that:

$$dsp(h, \xi) + dsp(\neg h, \xi) \leq 1 \tag{2.14}$$

and so that:

$$dsp(h, \xi) \leq pla(h, \xi) \tag{2.15}$$

In our work on this text we will deal with ETRI systems where the plausibility of our hypotheses will always be 1, and our focus will be on dsp values.

2.1.5 PAS Example

We finalize our treatment of PAS by presenting a probabilistic evaluation of our earlier example from section 2.1.2. Let us assume now that we make the following probability assignment Π for our assumptions:

$$p(a_1) = 0.5,\ p(a_2) = 0.3,\ p(a_3) = 0.2,\ p(a_4) = 0.1$$

This creates the previous table, this time enhanced with probabilities (Table 2.3)

Table 2.3. Scenarios for the example PAS instance

#	a_1	$p(a_1)$	a_2	$p(a_2)$	a_3	$p(a_3)$	a_4	$p(a_4)$	$p(s)$	$h=v_1$	$s \in QS_A(h, \xi)$	$s \in SP_A(h, \xi)$	$s \in QS_A(\perp, \xi)$
s_1	0	0.5	0	0.7	0	0.8	0	0.9	0.252	0			
s_2	0	0.5	0	0.7	0	0.8	1	0.1	0.028	0			
s_3	0	0.5	0	0.7	1	0.2	0	0.9	0.063	0			
s_4	0	0.5	0	0.7	1	0.2	1	0.1	0.007	0			
s_5	0	0.5	1	0.3	0	0.8	0	0.9	0.108	0			
s_6	0	0.5	1	0.3	0	0.8	1	0.1	0.012	0			
s_7	0	0.5	1	0.3	1	0.2	0	0.9	0.027	1	X	X	
s_8	0	0.5	1	0.3	1	0.2	1	0.1	0.003	$\perp$	X		X
s_9	1	0.5	0	0.7	0	0.8	0	0.9	0.252	1	X	X	
s_{10}	1	0.5	0	0.7	0	0.8	1	0.1	0.028	$\perp$	X		X
s_{11}	1	0.5	0	0.7	1	0.2	0	0.9	0.063	1	X	X	
s_{12}	1	0.5	0	0.7	1	0.2	1	0.1	0.007	$\perp$	X		X
s_{13}	1	0.5	1	0.3	0	0.8	0	0.9	0.108	1	X	X	
s_{14}	1	0.5	1	0.3	0	0.8	1	0.1	0.012	$\perp$	X		X
s_{15}	1	0.5	1	0.3	1	0.2	0	0.9	0.027	1	X	X	
s_{16}	1	0.5	1	0.3	1	0.2	1	0.1	0.003	$\perp$	X		X

Let $s = (x_1, \ldots x_m)$ be a scenario where x_i is the truth value of an assumption (0 or 1). Then the probabilities of scenarios are computed using:

$$p(s) = \prod_{i=1}^{m} p(a_i)^{x_i} \cdot (1 - p(a_i))^{(1-x_i)} \tag{2.16}$$

The computations for our example can be seen on column $p(s)$ on Table 2.3. We see that, the rows s_7 to s_{16} are part of the quasi-support for our hypothesis. To get the degree of quasi-support $dqs(v_1, \xi)$ we simply add the corresponding probabilities for those scenarios:

$$dqs(v_1, \xi) = \sum_{i=7}^{16} p(s_i)$$
$$= 0.027 + 0.003 + 0.252 + 0.028 + 0.063 + 0.007 + 0.108 + 0.012 + 0.027 + 0.003$$
$$= 0.53$$

For computing $dsp(v_1, \xi)$ we have to find $dqs(\perp, \xi)$ to be able to normalize the degree of quasi-support. This is present in the right-most column (s_8, s_{10}, s_{12}, s_{14}, s_{16}):

$$dqs(\perp, \xi) = 0.003 + 0.028 + 0.007 + 0.012 + 0.003$$
$$= 0.053$$

Summing the values on column for $SP_A(h, \xi)$ (s_7, s_9, s_{11}, s_{13}, s_{15}) and normalizing them we get:

$$dsp(v_1, \xi) = \frac{0.027 + 0.252 + 0.063 + 0.108 + 0.027}{1 - dqs(\perp, \xi)}$$
$$= \frac{0.477}{1 - 0.053} = \frac{0.477}{0.947} \cong 0.504$$

Using Definition 2.6 we also get the same result:

$$dsp(v_1, \xi) = \frac{p(QS(h, \xi)) - p(QS(\perp, \xi))}{1 - p(QS(\perp, \xi))}$$
$$= \frac{dqs(h, \xi) - dqs(\perp, \xi)}{1 - dqs(\perp, \xi)}$$
$$= \frac{0.53 - 0.053}{1 - 0.053} = \frac{0.477}{0.947} \cong 0.504$$

Now let us see how this result is reached using arguments instead of scenarios. For adding the probabilities as we did with scenarios, we have to ensure that they are disjoint sets. This is a classical problem in probability theory, and there is a range of algorithms to deal with them. It actually is known to be NP-hard and related to the famous satisfiability

(SAT) problem (Antoine *et al.*, 2003). The most basic of these methods is known as inclusion-exclusion principle, see (Antoine *et al.*, 2003) for a quick review. Here, we deal with the problem using manipulation on the logical equations.

$$\begin{aligned} dqs(\bot,\xi) &= p(a_1 \vee (a_2 \wedge a_3)) \\ &= p(a_1 \vee (a_2 \wedge a_3 \wedge \neg a_1)) \end{aligned}$$

Now we have made the two arguments disjoint, we proceed further;

$$\begin{aligned} dqs(\bot,\xi) &= p(a_1) + p(a_2 \wedge a_3 \wedge \neg a_1) \\ &= 0.5 + 0.3 \cdot 0.2 \cdot (1 - 0.5) \\ &= 0.53 \end{aligned}$$

This shows well that instead of dealing with an exponentially increasing number of scenarios, arguments serve our purpose better as long as it is possible to separate them into disjoint sets in an efficient way. Note that inclusion-exclusion principle similarly yields an exponential number of terms, so it is not a replacement in that sense. A commonly used algorithm for creating disjoint terms is Heidthmann's KDH algorithm (Heidtmann, 1989). We use binary decision diagrams as a way of coping with this complexity as shown in Appendix B.

Using similar techniques we get

$$\begin{aligned} dqs(\neg v_1,\xi) &= p(a_4) = 0.1 \\ dsp(\neg v_1,\xi) &\cong 0.05 \end{aligned}$$

and,

$$\begin{aligned} pla(v_1, \xi) &= 1 - dsp(\neg v_1,\xi) \\ &= 1 - 0.05 = 0.95 \end{aligned}$$

This relates to the amount that our hypothesis contradicts with our system. Thus we get that, our hypothesis has a degree of support 0.504 and a plausibility 0.95, the gap in between these figures represents our ignorance.

2.1.6 Applying PAS to Information Retrieval (IR): Enhancing relevance

Application of PAS to IR is a topic first explored by Picard (Picard, 1998). Of the many, we consider two aspects of interest in his application; enhancing relevance and computing a popularity measure. In his work the first case of enhancing relevance is widely explored and experimented, while the second one for computing popularities is not. This second topic will actually be a main focus of interest in this text.

We will briefly review these topics in this and the following sub-sections, the interested reader can consult (Picard, 1998), (Picard, 2000) and (Picard and Savoy, 2003) for further information. Also, we will effectively re-introduce and cover these models within our ETRI framework in chapters 2 and 3.

In his treatment of enhancing relevance, the author firstly considers spreading activation as an established competing method. In this method, an initial retrieval status value (RSV) is assigned for each document based on similarity for a query, and these values are spread to neighboring documents which are linked by hypertextual links.

$$RSV(d^0) = initial\_score(\, d,\, q\,) \tag{2.17}$$

$$RSV(d^{i+1}) = RSV(d^i) + \sum_{j=1}^{m} \lambda_j RSV(d_j^i) \tag{2.18}$$

where d is the document, q represents a query, i the count of iterations run, and λ_j ($0 < \lambda_j < 1$) is a parameter adjusting what fraction of the results will be propagated from a document to its neighboring documents.

The number of iterations this process is run (i) can be problematic because, it is claimed that running more than one iteration may harm the retrieval effectiveness.

The PAS solution is conceived in a similar manner. Instead of using an adhoc method of propagating values, PAS offers a clean and principled way of achieving a similar end not by propagation but by finding supporting arguments for a document's relevance, and using them to enhance values.

For this purpose a PAS knowledge-base is created using the document collection. Firstly, each document d_i receives a proposition v_i signifying its relevance stating "document d_i is relevant". For each such proposition a corresponding assumption variable a_i is associated such that:

$$a_i \rightarrow v_i$$

If the retrieval system assigns an initial relevance value to this document (retrieves it), then $p(a_i) > 0$. These initial values can be assigned directly by the retrieval system, or may be cast using logistical regression (Picard, 1998).

Secondly, the hypertext structure of the collection is reflected in the PAS knowledge. For each hypertextual link from a document d_i to a document d_j we gather that, if d_i is relevant than so must be d_j. In PAS we can encode this using:

$$v_i \wedge l_{ij} \rightarrow v_j$$

where l_{ij} is the "link" assumption from d_i to d_j denoting the condition under which this link implies relevance.

This rule can be read as: "If document d_i is relevant, then, under some condition l_{ij} (that the link from d_i to d_j implies d_j's relevance), d_j is also relevant."

Using these two constructs the whole PAS knowledge base is built. It is pre-processed once, and the supporting arguments are found for each document. Then, disjoint sets for computing degrees of support are computed and stored for each document.

The value assignments for link assumption probabilities $p(l_{ij})$ are explored in (Picard, 1998). The only option considered is assigning a constant value. Simply put, the average ratio of relevant documents out of the neighboring documents of a document is used as an estimate for this value.

Thus when the initial relevance values are assigned by a retrieval system, they can be enhanced using the pre-computed degrees of support formulations for the involved documents.

Note that both forward links and backward links represent evidence of relevance, and each of them can be used for this purpose. Since PAS formulation inherently deals with cycles without problem, both can be used at the same time as well. In contrast this would pose a problem spreading activation.

For example, let us assume backward links are used, and we have the following knowledge base:

$$\xi = (a_1 \rightarrow v_1) \wedge (a_2 \rightarrow v_2) \wedge (a_3 \rightarrow v_3) \wedge (v_2 \wedge l_{12} \rightarrow v_1) \wedge (v_3 \wedge l_{31} \rightarrow v_1)$$

The support for document d_1 can then be inferred using the knowledgebase:

$$SP(v_1, \xi) = a_1 \vee (a_2 \wedge l_{21}) \vee (a_3 \wedge l_{31})$$

The degree of support can then be computed by creating disjoint sets and adding their probabilities:

$$dsp(v_1, \xi) = p(SP(v_1, \xi))$$

$$= p(a_1) + p(a_2 \wedge l_{21} \wedge \neg a_1) + p(a_3 \wedge l_{31} \wedge \neg a_1 \wedge \neg (a_2 \wedge l_{21}))$$

$$= p(a_1) + p(a_2)\, p(l_{21})\, (1 - p(a_1)) +$$
$$p(a_3)\, p(l_{31})\, (1 - p(a_1))\, (1 - p(a_2)\, p(l_{21}))$$

In our experiments throughout this text we will stick to a forward links model, but our results are readily applicable for a backwards link model as well. However, as we will explore in chapter 4, our simplest model named ESP-0 is not capable of dealing with short-cycles properly, and it would reduce its effectiveness to use both backward and forward links at the same time.

2.1.7 Measuring Popularity with PAS

The authors Picard and Savoy explain briefly in (Picard and Savoy, 2003) how their PAS model may be used as a popularity measure. This will be a starting point for our work and this topic will constitute a major focus for us.

The PAS construction is very similar to the one in section 2.1.6. This time, the proposition p_i assigned to each document is taken to mean "document d_i corresponds to the user's interest".

It is assumed that an external source provides the initial values for the corresponding assumption probabilities $p(a_i)$. This may be gathered using a profile constructed from bookmarks, tracked from user's surfing pattern or specified explicitly using keywords by the user. Then we have in the PAS knowledgebase:

$$a_i \rightarrow p_i$$

Similarly the relevance measure similar to the previous sub-section, is taken to mean similarity in the user's interest, such that:

$$p_i \wedge l_{ij} \rightarrow p_j$$

The authors state that this way a personalized ranking scheme can be constructed, and that if each page is assigned an equal probability than it corresponds to PageRank (see section 2.3).

In our work we will focus on this non-personalized ranking scheme mentioned, explore its theoretical foundation, create efficient approximation methods, and try to demonstrate that it actually creates a powerful ranking scheme competing with PageRank.

2.2. Complex Networks

Our work is an attempt towards adding a new tool of research for complex networks. In that, as an initial and primal way of using it we have dealt with link analysis ranking. Yet we believe that our work can be put to good use also for other network analysis means like community/topic detection, examination of components/hierarchies amongst others. In chapter 6, we examine the small-world properties of our experimental network and also investigate complex network properties of the ranking algorithms we have introduced.

Interest in complex networks has seen a recent increase, and many properties have been studied. This is partly because networks of different kinds like World Wide Web have started to play an important role in human life. Earlier interest had been mostly on random graphs also referred to as Erdös-Renyi graph models, but more recent work has focused on complex/social networks (Newman, 2003) or the “small-world model”.

Social networks such as the web, internet, affiliation networks and many others have been shown to share some interesting “small-world” properties, which we deal shortly in the next section. In section 2.2.2 we will present Zipf plot which has been a useful tool for us in identifying power-law distribution exponents for the networks we have examined.

2.2.1 Small-world Network Model

Watts and Strogatz have initially introduced the small-world network model (Watts and Strogatz, 1998). The term is mainly meant to refer to three characteristics of complex networks. These are:

i. Average path length

The geodesic average path length of small-world networks have been shown to be unexpectedly short, in comparison to purely random networks. This was firstly exemplified in Milgram's classic work (Milgram, 1967). For example, a 200 million node snap-shot of WWW has been shown to have an average path length of 16.18 (Newman, 2003), where as Milgram had found an average separation of 6.

ii. Clustering coefficient

A clear deviation from random behavior has been observed in complex networks towards clustering. That is, if a vertex *A* has a link with *B*, and *B* with *C*, then *A* has a heightened probability to have a link with *C*. There are two versions of clustering coefficients offered in the literature both of which work to present a measure of this behavior. In (Watts and Strogatz, 1998), authors present a re-wiring based generative model to create a network which desired clustering coefficients. The two formulas are:

$$C_1 = 3 \text{ x } \textit{number of triangles in the network / number of connected triples of vertices} \tag{2.19}$$

$$C_2 = 6 \text{ x } \textit{number of triangles in the network / number of paths of length } 2 \tag{2.20}$$

iii. Degree distributions

It has been observed that many complex networks exhibit a skewed degree distribution, and that most of them can best be described using a power-law distribution. That is, for some characteristic exponent α the degrees follow:

$$N(x) \propto x^{-\alpha} \tag{2.21}$$

where $N(x)$ can be in-degree (citation count), or out-degree.

This property has been termed "scale-free". It has also been shown that other vertex properties like PageRank also follow this distribution (Pandurangan *et al.*, 2002). We detail more on this in section 2.3. Some of our introduced measures have also shown this distribution as shown in section 5.7.

An interesting point is that, two networks of high interest have different characteristic exponents. For the web graph, the value is ~ 2.1 (Newman, 2003), whilst for scientific citation networks it is rather ~ 3.0 (Redner, 1998)(Redner, 2004). Our findings on our examined citation network confirm the value 3.0.

2.2.2 Zipf-plot

Zipf-plot is a valuable tool firstly introduced in (Zipf, 1949). We follow the example in (Redner, 1998) to determine the power-law exponents in our experiments. Quoting from (Redner, 1998):

"To help expose the differences in the citation distribution, it is useful to construct a Zipf plot (Zipf, 1949), in which the number of citations of the k^{th} most-ranked paper out of an ensemble of M papers is plotted versus rank k (as in Figure 6.8). By its very definition (see Eq.(2.22)), the Zipf plot is closely related to the cumulative large-x tail of the citation distribution. This plot is therefore well-suited for determining the large-x tail of the citation

distribution. The integral nature of the Zipf plot also smooths the fluctuations in the high-citation tail and thus facilitates quantitative analysis.

Given an ensemble of M publications and the corresponding number of citations for each of these papers in rank order, $Y_1 \geq Y_2 \geq \ldots \geq Y_M$, then the number of citations of the k^{th} most-cited paper, Y_k, may be estimated by the criterion:

$$\int_{Y_k}^{\infty} N(x)dx = k \tag{2.22}$$

where $N(x)$ is the number of papers with x citations.

This specifies that there are k publications out of the ensemble of M which are cited at least Y_k times. Eq.(2.22) also represents a one-to-one correspondence between the Zipf plot and the citation distribution. From the dependence of Y_k on k in a Zipf plot, one can test whether it accords with a hypothesized form for $N(x)$."

Similar to citation count, PageRank and our introduced ranking algorithms have large and fluctuating tails. Therefore we have used Zipf plots successfully to determine their power-law exponents as well.

It is straightforward to determine the power-law exponent after fitting a line to the Zipf plot. For a power-law distribution of the form in Eq.(2.21), Eq.(2.22) gives for large x:

$$\alpha = 1 - \frac{1}{b} \tag{2.23}$$

where b is the slope of a fitted line (see Figure 6.8, Figure 6.11, Figure 6.12 as examples).

2.3. Link Analysis Ranking (LAR) Algorithms

In 1998, two influential papers have created a new research area which might be termed as "link analysis ranking" (Langville and Meyer, 2004) (Borodin *et al.*, 2005). These introduced the PageRank ranking algorithm (Page *et al.*, 1998), and HITS algorithm (Kleinberg, 1999).

A great deal of effort has been made on analysis of these algorithms, and extensions and enhancements have been suggested. For our part, we will focus on PageRank because it has been part of our inspiration for our algorithms along with PAS based ranking (Picard and Savoy, 2003).

A good general treatment of ranking algorithms with experimental evaluations using human testers can be found in (Borodin *et al.*, 2005). Another evaluation of PageRank along with a PAS based ranking algorithm is presented in (Savoy and Rasolofo, 2000).

In the sub-sections to follow we present an introduction to the PageRank algorithm, and later on we review some of its critical evaluations.

2.3.1 PageRank Algorithm

PageRank has been studied extensively by numerous authors, certainly in part due to its impact on the internet experience being used in the Google search engine. There has been interest in its mathematical foundations (Langville and Meyer, 2004), its efficient application (Haveliwala, 1999), and its approximations (Chen *et al.*, 2004) amongst others.

With more direct implication for our work are the ones which propose extensions to the basic algorithm, because most of these are readily applicable to our algorithms as well (Richardson and Domingos, 2001) (Ingongngam and Rungsawang, 2003) (Haveliwala, 2002) (Kao *et al.*, 2002). These provide a natural future direction of extension for our work.

PageRank is conceived as an extension to citation counting, in which the significance of the citing document is also taken into account. This way it becomes a global measure of importance, and it is considered to contain more information. A well quoted example is that, the main page of Yahoo! search engine has more significance than an ordinary page. So, if a web-page receives a citation from the Yahoo main page, that citation should have more significance.

PageRank is presented in two different formulations. The first one, termed as the simple (iterative) formulation is:

$$Rank_{i+1}(v) = \sum_{u \in B_v} \frac{Rank_i\,(u)}{N_u} \tag{2.24}$$

where B_v is the set containing the parents of a vertex v, and N_u is the number of links going out of document u (out-degree).

While presenting the idea clearly, this fails to be applicable for the web because in this model ranks can get trapped in an isolated cluster of the graph, in which two pages only link to each other acting as a “rank sink”.

The second formulation addresses this problem, by adding a rank source, and discounting for the additional source using a damping factor d.

$$Rank_{i+1}(v) = \left(\frac{1-d}{n}\right) + d \sum_{u \in B_v} \frac{Rank_i\,(u)}{N_u} \tag{2.25}$$

Here n is the number of vertices in the network. It is assumed that, either all vertices with 0 out-degrees are iteratively pruned (and added back after ranking is done), or that virtual links going out from such vertices to all the vertices in the network are added to the network.

There are a number of additional ways of interpreting this formulation. One is the random surfer model. In this, a random surfer is assumed to be surfing the web following random outgoing links on a page with probability *d*, and making a completely random jump to a page in the web with probability 1-*d*. In this sense, the PageRank vector is the stationary probability of a random walk on a Markov chain created using the "Web graph".

It is also possible to view it as the primary eigenvector of the created transition matrix which has been made stochastic and irreducible thus ensuring the existence of a stable eigen-vector. For example, in (Jeh and Widom, 2003) the authors take this vector interpretation further ahead. These and other important mathematical details like convergence are deeply explored in (Langville and Meyer, 2004).

We have seen variations of the formula in 1.12 in the literature. The denominator *n* in the first half of the equation is missing in some papers on PageRank. Confusingly, it appears to be given wrong in the initial technical report which introduced PageRank! (Page *et al.*, 1998).

2.3.2 Usefulness of PageRank and PageRank vs. Citation Count

The usefulness of PageRank appears to be doubted in the literature. On the one hand, the original authors report impressive enhancements (Page *et al.*, 1998)(Brin and Page, 1998) coupled with the commercial success of the Google search engine which reportedly uses it.

There are findings which claim that PageRank experimentally performs worse than simple citation counting (Borodin *et al.*, 2005). Some authors claim that PageRank's of pages are highly related to citation counts (Ding *et al.*, 2002) (Upstill *et al.*, 2003), while others dispute that (Pandurangan *et al.*, 2002).

Authors in TREC conferences have not found any improvements in using PageRank (Ingongngam and Rungsawang, 2003) (Savoy and Rasolofo, 2000)(Picard and Savoy,

2003) or citation based algorithms (Savoy and Picard, 1999) over content based ranking schemes.

It appears that whilst there is consensus in the usefulness of employing link based information, more research is necessary in this area. The way of using link information along with the rank merging problem (combining different sources of ranking) may surface to be foremost issues to be addressed.

We will be applying PageRank to a scientific citation network in our work. Its usefulness in this sense might be doubted as the random surfer intuition does not apply as readily to the scientific research process. Yet still, there is some truth in this model even for scientific research, and even if not, the initial justification for extending simple citation counts to a more global ranking scheme still fully apply.

2.4. Mathematical Background for Our Models

In this section we explore the background for, and introduce an important mathematical operator we use extensively in the rest of the text. It is not necessary for following the text (we also give the equations without using it), but we have found it very useful in assisting our proofs and shortening our equations ultimately making them much more intuition friendly.

2.4.1 Sylvester-Poincare Formula for Pair-wise Disjoint Terms

We review here a useful application of the Sylvester-Poincare (Inclusion-Exclusion) formula for disjunction of disjoint terms. Consider the following situation for computing the probability for the disjunction $T_1 \vee T_2$ of two terms of arbitrary order (num. of literals) T_1 and T_2, which we know, are pair-wise disjoint. Using the Sylvester-Poincare development we get:

$$
\begin{aligned}
p(T_1 \vee T_2) &= p(T_1) + p(T_2) - p(T_1 T_2) \\
&= p(T_1) + p(T_2) - p(T_1)\, p(T_2) \\
&= 1 - (1 - p(T_1))\,(1 - p(T_2)) \qquad (2.26)
\end{aligned}
$$

We illustrate the idea below that, using incremental application of Eq.(2.26) along with associativity and commutativity of disjunction, it can be shown that the probability of the disjunction with additional pair-wise disjoint terms T_3, T_4, ... T_n creating $T_1 \vee T_2 \vee T_3 \vee T_4 ... \vee T_n$ is:

$$
\begin{aligned}
p(\bigvee_{i=1}^{n} T_i) &= 1 - (1 - p(T_n)) \left(1 - p(\bigvee_{i=1}^{n-1} T_i)\right) \\
&= 1 - (1 - p(T_n)) \left[1 - \left(1 - \prod_{i=1}^{n-1} (1 - p(T_i))\right)\right] \\
&= 1 - (1 - p(T_n)) \left(\prod_{i=1}^{n-1} (1 - p(T_i))\right) \\
&= 1 - \prod_{i=1}^{n} (1 - p(T_i)) \qquad (2.27)
\end{aligned}
$$

Note that, this actually is the formulation for the noisy-or gate (Heckerman and Breese, 1996).

2.4.2 Noisy-or Operator

In this section we introduce the ***noisy-or operator*** "$\hat{\vee}$" which give us convenience for expressing certain types of mathematical formulations in the following sections.

Definition 2.7. (**noisy-or operator**) *For* $a,b \in \mathbb{R}$ *and* $0 \le a,b \le 1$ *we define the binary operator "* $\hat{\vee}$ *" in the infix form such that:*

$$a \,\hat{\vee}\, b = 1-(1-a)(1-b) \tag{2.28}$$

Theorem 2.1. *Noisy-or operator has the following properties:*

(1) Commutativity
(2) Associativity

Definition 2.8 (noisy-or operator / prefix form) *We define a pre-fix form of the noisy-or operator "* $\hat{\vee}$ *" such that*

$$\hat{\bigvee_{i=1..n}} a_i = a_1 \,\hat{\vee}\, a_2 \,\hat{\vee} \ldots \hat{\vee}\, a_n \tag{2.29}$$

where $0 \le a_i \le 1,\ 1 \le i \le n$*, and n,i are positive integers.*

Definition 2.9 (precedence of noisy-or) *We define the precedence of the noisy-or operator such that it has higher priority than addition/subtraction, and lower priority than multiplication/division.*

Example: $a \cdot b \,\hat{\vee}\, c + d = \left((a \cdot b) \,\hat{\vee}\, c\right) + d$

Note that;

$$\hat{\bigvee}_{i=1..n} a_i = a_1 \hat{\vee} a_2 \hat{\vee} ... \hat{\vee} a_n = 1 - \prod_{i=1}^{n} (1 - a_i) \tag{2.30}$$

We can thus re-write the probability of the propositional sentence of Eq.(2.27) $T_1 \vee T_2 \vee T_3 \vee T_4 ... \vee T_n$ with pair-wise disjoint terms using the noisy-or operator:

$$p(\bigvee_{i=1}^{n} T_i) = \hat{\bigvee}_{i=1..n} a_i \tag{2.31}$$

Theorem 2.2 *Let a, b* $\in \mathbb{R}$ *be such that* $0 \leq a, b \leq 1$. *Then the following equality using the noisy-or operator holds:*

$$a \hat{\vee} b \geq a \tag{2.32}$$

3. PAS ENTITY-TRANSITIVE RELATION-IMPLICATION (ETRI) MODEL

3.1. Introducing the PAS-ETRI Model

In this section we define a graphical model for describing transitive relations between different entities of a domain.

Definition 3.1 (**PAS-ETRI Model**)

A PAS-ETRI is a tuple $ETRI(G, PAS_{ETRI}, R)$ *where G is a directed graph* $G = (V, E)$, PAS_{ETRI} *is a type of PAS such that* $PAS_{ETRI} = (\xi, P, A, \Pi)$, *R is a semantic transitive relation. We further specify the following:*

Let n be the number of vertices, m be the number of arcs (directed edges) in G, we specify V, P, and E as:

$$V = \{v_1, v_2, ..., v_n\} = P \tag{3.1}$$

$$E = \{e_1, e_2, ..., e_m\} \tag{3.2}$$

We will refer to the elements of sets V = P as ***entities***.

The assumptions A in PAS_{ETRI} *are defined as:*

$$A = \{a_1, a_2, ..., a_n\} \cup \{l_{ij} : 1 \le i, j \le n, i \ne j, 1 \le k \le m, \text{ such that there exists an arc } e_k \text{ from vertex } v_i \text{ to } v_j \text{ in } G\} \tag{3.3}$$

We will refer to the subset $a_i \in A$ *as* ***node assumptions****, and the subset* $l_{ij} \in A$ *as* ***link assumptions****.*

The knowledge-base ξ *is specified as:*

$$\xi = \{a_i \rightarrow v_i : a_i \in A\} \cup \{v_i \wedge l_{ij} \rightarrow v_j : l_{ij} \in A\} \tag{3.4}$$

For the semantic transitive relation R we specify:

> *Let* $a,b,c \in D$ *be entities of a domain D where R holds. Then the following should be correct:*

$$\forall a,b,c \in D : \big(R(a,b) \wedge R(b,c)\big) \rightarrow R(a,c) \tag{3.5}$$

The PAS-ETRI model is built on the work Picard to apply PAS for Information Retrieval which we survey in sections 2.1.5 and 2.1.6 (Picard, 1998) (Picard, 2000) (Picard and Savoy, 2003). Here, we essentially formalize and generalize the "hypertext retrieval model" presented in (Picard, 2000), along the lines of (Picard and Savoy, 2003) for a general entity-relation setting.

PAS-ETRI model corresponds to a particular type of PAS where the knowledge base is made of horn clauses. The propositional sentences inferred using this knowledge-base have the following basic form:

$$\big(a \wedge b \wedge ... \wedge\big) \vee \big(c \wedge d \wedge ... \wedge\big) \vee ... \vee \big(e \wedge f \wedge ... \wedge\big) \rightarrow g \tag{3.6}$$

It is known that deciding entailment for a proposition in such a knowledge base can be done in linear time in the size of the knowledge-base (Russell and Norvig, 2003). Effectively, this kind of inference can be perceived as a path finding process in a graph.

Theorem 3.1. (Support for a Vertex)

Given an $ETRI(G, PAS_{ETRI}, R)$, *the support* $SP_{ETRI}(h, \xi)$ *for* $h = v_i$ *is:*

$$SP_{ETRI}(v_i, \xi) = a_i \vee \bigvee_{j \in P_i} \left[l_{ji} \wedge SP_{ETRI}(v_j, \xi) \right] \tag{3.7}$$

This is identical to saying that supporting arguments for a vertex contain all of its parents' arguments (conjuncting with relating link assumptions) in addition to its own assumption. This formulation is shown for hypertext retrieval model in (Picard, 2000).

For listing supporting arguments, we will use the set representation (as a collection of terms/scenarios) and sentence representation (as a DNF sentence) for $SP_{ETRI}(v_i, \xi)$ interchangeably.

This kind of knowledge base contains no contradictions. Computing degrees of support for vertices is easier as the quasi-support for an hypothesis is equal to its support, and as pointed out in (Haenni, 2003) there exists many efficient satisfiability (SAT) problem based methods for dealing with such knowledge bases. In Appendix B, we introduce our implementation of a PAS-ETRI using Binary Decision Diagrams (BDD) (Bryant, 1986), which form the basis of some such efficient methods.

Intuitively, a PAS-ETRI is a semantic network containing one type of link, backed by PAS, and containing associated probabilities for nodes and links representing their significance. PAS are a special-case of Dempster-Shafer theory of evidence (Haenni *et al.*, 2000). Thus it is natural to expect that "significance" is the belief in some evidence which assumes its true meaning depending on what the semantic relation contained is. In the following section, we try to explore a variety of such systems to elaborate on the usefulness of the model introduced.

Example 3.1.

Consider a PAS instance where the knowledge-base is:

$$\xi = \{\ a_1 \rightarrow v_1\ ,\ a_2 \rightarrow v_2\ ,\ v_2 \rightarrow (a_3 \rightarrow v_1)\ \}$$

This is essentially our previous PAS example with the last clause removed. Here we can see that the knowledge base has no contradictions, and is made of only Horn clauses. We see that for $h = v_1$ the support and the quasi-support are the same:

$$QS(h, \xi) = SP(h, \xi) = a_1 \vee (a_2 \wedge a_3)$$

See Figure 3.1 for a graphic representation. Here the literals in squares are the assumptions, and the literals in the circles are the propositions. We can also see how a graphical structure is mapped to a PAS instances, replacing implication "arrows" in the knowledgebase with graph "arrows".

We can see how finding the support corresponds to walking backwards from the hypothesis node v_1, to the "supporting" nodes.

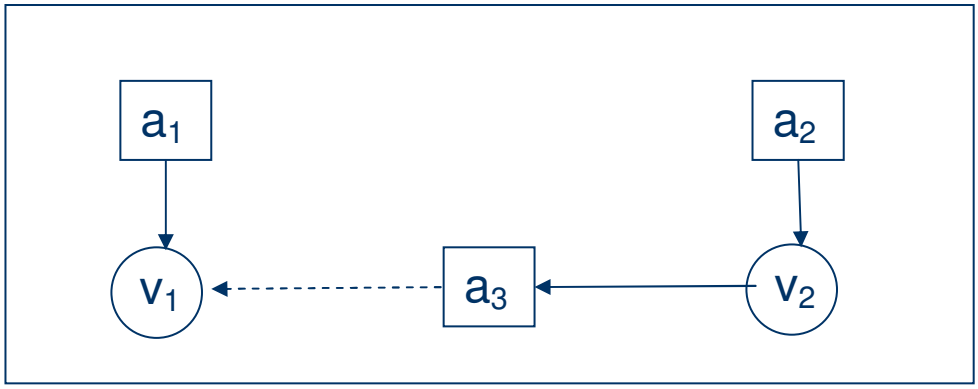

Figure 3.1. Example PAS-ETRI network

3.2. Possible Applications of the PAS-ETRI Model

The PAS-ETRI model can find application in a variety of different domains where a network based modeling has made sense. We perceive it as a tool for analyzing complex networks of all kinds.

Many different kinds of such networks can be named; different kinds of complex networks (www, citation networks, friendship/acquaintance networks, networks for spreading of diseases, ...), trust networks, different social networks such as organizational hierarchies, information retrieval networks (hypertext retrieval), software function call graphs, biological networks (e.g. neural networks). This list can be further extended, instead we will focus on a few models to illustrate the motivation for the model.

When PAS-ETRI is applied for Information Retrieval (IR) (see section 2.1.5 and 2.1.6), the semantic relation contained can be "relevance". Picard defines relevance using the concept of "infons" (Picard, 2000). Infons are defined to be elementary items of information individuated by a cognitive agent. The probabilities of the node assumptions then represent our prior "evidence" that a document is relevant (i.e. contains infons that are relevant given a query), whilst the link assumptions can be thought of as further evidence as for which this relevance can be expanded, thus resulting in posterior probabilities of relevance, namely "degree of support".

As pointed out in (Page *et al.*, 1998) recall is not the only problem for IR in large collections. The ranking of documents becomes a focal point when there are many equally relevant documents. PAS-ETRI forms naturally a tool for link based ranking of documents in a collection as suggested in (Picard and Savoy, 2003) (surveyed in section 2.1.6). The solution of the ranking problem necessitates another ETRI model. In chapter 3, we focus on this problem, and introduce a model centered on the concept of "information value" instead of relevance.

For analyzing community structures for authors of scientific papers, a PAS-ETRI can be constructed using a relation "influences". Obviously, this is not a strictly transitive

relation. Yet, useful analysis can still be made, keeping this fact in mind and interpreting the results with according reservation. In such a setting, a PAS-ETRI system can be asked to identify the most "influential" author, or with additional modeling (e.g. using disjunctions of vertices) an attempt to reveal community structures may be made.

For social networks, the emphasis may be on different kinds of relations. One such example is spreading of diseases. A virtually transitive relation "infects" can be used. Note that, when using PAS-ETRI, detailed and more precise modeling of interactions between each individual can be specified. The inability of some models to do this has previously been criticized (Handcock *et al.*, 2003). In such a system, one could ask to identify the most "infectious" individuals. As such, possible changes to the network such as effects of vaccination can be investigated under clear semantics.

It is important here to point out that, all the results derived from PAS are derived under strong independence assumptions. That is, all link and node assumptions are assumed to be stochastically independent. Obviously, for the disease setting above like some others, it is a reasonable assumption to make that the infections of two completely stranger individuals would be independent. Yet we believe that the way of modeling a particular phenomenon would benefit from paying attention to expressing the system in a way that maximizes the correspondence of independence assumptions to the modeled system.

4. PAS-ETRI AS A LAR TOOL

4.1. Applying PAS-ETRI for Information Retrieval

As has been suggested in the previous chapter, a PAS-ETRI based analysis can be made in various ways even for the same problem. For the task of finding appropriate documents in a collection for a query, we will present two candidate models. The first one focuses on ranking the documents according to their "information value", while the second one focuses on finding the relevant documents. Combined together these two models are meant to return the "relevant" documents which have a "high information value", thus resulting in an effective method of retrieval. We essentially build on the models developed in (Picard and Savoy, 2003) and (Picard, 1998) (presented in sections 2.1.6 and 2.1.5) using the introduced ETRI framework.

This way of combining relevance and value of information content has been suggested and applied successfully in a slightly different context in (Page *et al.*, 1998). As was reviewed in section 2.3, it was suggested to rank documents according a popularity measure based on random walks on the graph. In our work we differ by replacing the concept of popularity with a new one; "information value" which we will define in the following sections. Our definition of information value will more closely follow "citation count" which predates PageRank as possibly being the earliest measure of importance.

In this chapter we introduce ArgRank, a novel ranking algorithm, which we build on a concept of Minimal Evidence (ME) on the PAS-ETRI model. For the ranking approach each document receives a value representing its information value and this represents the "link evidence" gathered from the network. This approach of using PAS is firstly presented in (Picard and Savoy, 2003), but the idea is not fully explored. Thus, in this chapter we try to develop the idea in full detail.

In contrast, applying PAS for the relevance problem has been widely studied in (Savoy and Rasolofo, 2000) (Picard, 1998) as shown in chapter 2, from which we borrow

many ideas. In these papers, the authors have demonstrated the use of argumentation for "spreading" an initial relevance of documents to neighboring documents, thus enhancing the results.

In our work we will use a very simplified model for assessing relevance and focus on the ranking model. Once the ranks are assigned, a simple keyword match will be used to filter out documents, and this will be the way to combine two different sorts of evidence (rank merging) for relevance and information value. An analogous to this approach was suggested in (Page *et al.*, 1998).

4.2. ETRI models for Information Retrieval

In this part we will introduce the two IR related PAS-ETRI models which are designed to deal with the ranking and relevance assessment problems. In the ETRI, the vertices in the graph will represent documents (e.g. a web-page, a paper), and links are present whenever a documents cites or is cited by (or both) another document (depending on the way the graph is constructed).

For the first model, we define the transitive relation R to be "*is informative*". Then, a link from document *i* to *j* is taken as "*according to document i, document j is informative with probability* $p(l_{ij})$". The node assumptions are the a priori judgment that a document is informative. The term "informative" is used to judge the information quality of a document. In analogy with (Picard, 2000), it may be referred to as a measure for the amount of infons a document contains. This model mimics the PAS alternative to PageRank model presented in (Picard and Savoy, 2003) using PAS-ETRI terminology.

We will refer to this model as ***document/information value model*** **(DIM)**. We will use it to assess the quality of documents. Thus given a set of relevant documents, it will be possible to order and present them in decreasing quality.

The quality of a document in being informative, is not directly observable in an objectively measurable way, if not a speculative issue. Also, although a document's

references certainly signify that the author of the given document has found at least some of the referenced papers "informative", it remains a non-trivial issue to determine how the link assumptions should be assigned. In the rest of this work, we will try to address these problems.

Below we will introduce the PAS-ETRI model dealing with relevance. Although in the experimental part we will be using a simplified relevance model as mentioned above, we introduce this second model for the sake of completeness.

We define our transitive relation *R* as "is relevant to". So, a link from a document *i* to *j* is taken as "*according to document i, document j is relevant to document i with probability* $p(l_{ij})$" (reading backwards from the arrow direction). The node assumptions then represent our prior belief that a document is relevant. This may be supplied by a different IR system, or another source. This model is essentially developed and used in (Picard, 1998) and (Picard, 2000), and stated here using PAS-ETRI terminology.

We will refer to the PAS-ETRI model defined above, as ***document/relevance model* (DRM)**. Its counterpart is "hyper-text retrieval model" in (Picard, 2000). This model is useful for identifying documents that are relevant, given a query.

It should be noted that, for a given document the quality of being informative can be safely assumed to be independent of the topic of the document – although presumably there will be exceptions. This means that whether a document is relevant or not is not dependent on its information value. Thus as suggested in the previous section, the two models DRM and DIM are meant to be used in a complementary manner.

4.3. Minimal Evidence (ME)

We present here a node assumption probabilities assignment which we will use for facilitating discussions about merits of different link analysis ranking algorithms.

We define:

Definition 4.1 (**Minimal Evidence**) *For an ETRI*(*G, PAS, R*) *with n documents in the collection P, we define the partial assignment:*

$$p(a_i) = 1/n \quad where \quad p(a_i) \in \Pi \ , \qquad i = 1, 2, \ldots n \tag{4.1}$$

to be the Minimal Evidence (ME). Link probabilities are left unspecified in ME.

Intuitively, this way of assigning prior probabilities corresponds to the minimal *evidence* one has that, at least one document (and possibly, only one document) in the collection has the desired quality dictated by the transitive relation (e.g. is informative or is relevant).

Obviously, if we have prior knowledge that there is no document in the collection which is, say relevant, it does not make sense to look for it in the collection!

One may have noticed the similarity between the maximum-likelihood (ML) hypothesis for maximum a posteriori (MAP) learning and ME assignment. This actually is not a coincidence. In MAP learning, setting priors in this way corresponds to a distrust of priors, and it is useful for large data-sets where all hypotheses are equally complex and likely to be true (Russell and Norvig, 2003). Note the reminiscence of our setting with the situation described. It is our intention to "extract" the link based evidence on the graph, and this way of setting the prior assumptions thus allows us to focus on utilizing this information.

In the rest of the text, the soundness of the use of ME will become further clearer by the cases examined.

4.4. Introducing ETRI Ranking: ArgRank

In this section, we introduce a document ranking algorithm based on the ETRI model. We will refer to it shortly as **ArgRank**.

Definition 4.2 (ArgRank) *Given an ETRI*(*G, PAS, R*) *with the ME partial-assignment, we define the ArgRank of a document* v_i *as:*

$$ArgRank_i = dsp(v_i,\xi) = dsp_i \tag{4.2}$$

We have used the short notation dsp_i for the degree of support of a node. We will use this notation further on.

In the following sections we will use ArgRank to get a query and user independent ranking scheme that ranks the whole collection. Note that, while we have defined ArgRank using ME, it is possible to generate similar rankings for more "personalized" results by altering node assumption probabilities depending on the context (e.g. the inquirer) as in PageRank. So, we will sometimes use the term ArgRank to refer to such a family of rankings including the ones with altered (e.g. personalized) node assumptions.

A very similar ranking scheme based on PAS is suggested and examined in (Picard and Savoy, 2003) as an alternative to PageRank. However, the significance of ME is omitted and an emphasis is made on producing a personalized popularity measure using external evidence sources such as bookmarks.

4.5. Time-complexity Considerations for ArgRank Calculations

The computation of degree of support can have a high time-complexity if the number of arguments is high. In the PAS-ETRI model we need to find the disjoint terms of a DNF sentence for calculating the degrees of support. This problem is known to be NP-hard,

being related to counting all the solutions to the satisfiability (SAT) problem (Antoine *et al.*, 2003).

For many DRM settings, link assumptions are set to be relatively low (e.g. 0.05 to 0.3). Thus for the case of enhancing relevance, this results in not needing terms with more than a few literals in the arguments, as the marginal contribution of additional literals would not be worth paying the computation costs. Previous PAS related work includes projects with an order limit (max. number of literals in each supporting argument) of two (immediate neighbors) and three on the terms (Savoy and Rasolofo, 2000) (Picard, 1998). These authors report of no difficulties of calculations relating to collection size.

For the DIM case however, such an order limit may cause a high degradation on the accuracy of the results. In ArgRank, because of ME assignment the collective evidence provided by all the nodes is evenly spread in the collection. So in a larger collection with sufficient connectivity, this in effect may create concentrations of evidence for different "areas" of nodes. Rankings computed using a strict order limit would be less capable of reflecting this characteristic the bigger and more concentrated such areas are (e.g. a large group of documents with high citations referencing each other).

Also even when an order limit is imposed, while the question of time complexity may not be a significant issue for relatively smaller collections, the web is vast and its size is doubling in less than a year (Broder *et al.*, 2000a) (Page *et al.*, 1998). Any algorithm that has a time complexity significantly higher than linear amortized time would pose a very high challenge of application to scale to the web.

Following this discussion, it can be seen that an unlimited and straightforward application of ArgRank to a sizeable collection may be a very difficult if not an impossible task. To address this problem, in chapter 4 we explore a variety of methods including the order limit, and introduce a family of novel algorithms for approximating dsp values in an ETRI.

4.6. Comparing ArgRank and PageRank

PageRank is a link analysis ranking algorithm that certainly scales to the web currently being used in the famous Google search engine. This can be mainly attributed to the calculate-once nature (ignoring updates to the collection) for the initial ranking process. Only local computations (i.e. from immediate neighbors of a node in a graph) are used in a quickly converging scheme, yielding linear time-complexity in the number of edges.

While PageRank appears to be successful, it is unclear how the values produced by the algorithm should be interpreted from an AI or evidential reasoning perspective or should be combined with later evidence (e.g. rank merging). This may have contributed to the dispute on the usefulness of PageRank which we have presented section 2.3.2.

ArgRank on the other hand builds on evidential reasoning with clear semantics. The rankings obtained by ArgRank are, degrees of support defined within the PAS-ETRI model, which are in turn posterior probabilities for the relation (i.e. being informative) being true. Being a special case of Dempster & Shafer Theory, PAS effectively builds upon the theory of evidence. The ME closely mimics maximum likelihood (ML) hypothesis in MAP learning, a well-known method in statistics, for the evidence domain. Thus in effect, ArgRank is a result of combining some well established methods in a novel yet clearly demonstrable manner.

While the prospect of using ArgRank appears promising, it is set back by its reliance on requiring to perform computations for an NP-hard problem for each node in a vast collection. In the following part, we focus on tackling this problem for the general ETRI case, examining various solutions, and present novel ones.

5. EFFICIENT APPROXIMATE SOLUTIONS OF AN ETRI SYSTEM

5.1. An Assessment of Approximation Techniques

Following the discussion in section 4.5, it is quite obvious that an unlimited and direct application of ArgRank is unfeasible. As ArgRank is relying exclusively on NP-hard problem computations for a possibly vast collection of documents, we certainly do not have any good reason to expect that this situation should get better in the future.

Ruling out direct and unlimited application, this effectively makes the question of how to approximate ArgRank accurately and efficiently, a focal part of our work further on.

Reflecting on the discussion of chapter 4, we may formulate desirable characteristics of a link analysis ranking algorithm:

- Incorporate as much evidence from links as possible.
- Scale well to vast collections. Preferably require at most near linear time operations when dealing with the whole collection.
- Be theoretically sound.

These will also be our guidelines for devising and evaluating different approaches for the approximation methods.

In the following sections we explore and analyze various methods of approximating dsp values. Inspired by PageRank, we have a focus on methods using local computations (i.e. relating to immediate neighbors of a node).

The methods we will introduce for approximations are all applicable for dsp calculations for the general ETRI case independent of the model used. However, we will choose to focus on the DIM and the ArgRank for evaluations.

5.2. Imposing a Limit of 2nd Order for Supporting Arguments

An immediate candidate method fitting the criteria of the preceding section is the simplified ArgRank which imposes a 2-literal limit on arguments. As only the immediate neighbors are considered, the algorithm relies only on local-computations, and works in exactly linear time in the number of edges. This approximation for the ETRI-DRM context has been used in (Savoy and Rasolofo, 2000) previously.

It is interesting to note that in this model, for dsp calculations we effectively get the formulation for the noisy-or gate (Heckerman and Breese, 1996).

Using the ME assignment with a fixed link assumption probability p_l, and choosing to use "forward links" (i.e. link from v_i to v_j means "document *i has cited* document *j*") we get:

$$
\begin{aligned}
dsp_i &= 1-(1-p(a_i))\prod_{j\in P_i}\left(1-p(l_{ji})\,p(a_j)\right) \\
&= 1-(1-p_a)\prod_{j\in P_i}(1-\frac{p_l p_a}{n}) \\
&= 1-(1-p_a)\prod_{n=1}^{InDegree_i}(1-c) \\
&= 1-(1-p_a)(1-c)^{InDegree_i} \qquad (5.1)
\end{aligned}
$$

where $c=\frac{p_l p_a}{n}$, $p_a=\frac{1}{a_i}$, P_i represents parents of node *i*, $dsp_i = dsp(v_i,\xi)$.

Note that since we have used the "forward links", it follows that there is a link from node v_i to v_j if document *i* has cited *j*. So, the parents of a document are the documents

which cited it. Thus, multiplying *dsp*'s from all the citing documents has yielded $InDegree_i$ above.

We observe here that $c << 1$ for any sizeable collection. In this case, as long as $dsp_i << 1$, the following holds:

$$dsp_i = 1-(1-p_a)(1-c)^{InDegree_i} \cong p_a + c \cdot InDegree_i \propto InDegree_i \quad (5.2)$$

We identify Eq.(5.2) as possibly the oldest link analysis ranking algorithm, namely the ***citation count***. Note that p_a values are the same for all nodes, and that is how we can relate the ranks solely on in-degree.

The condition $dsp_i << 1$ holds as long as a document is not cited by a significant fraction of the collection, which is virtually impossible for most of the collections. This appears supportive for the soundness of our approach of using ME for evaluating link analysis ranking algorithms.

Note that, we have implicitly used the Boole-Bonferroni bounds in Eq.(5.2) for approximating the *dsp* value (Antoine *et al.*, 2003).

5.3. Total Independence Assumption for Supporting Arguments

Relating the dsp values of neighboring nodes is a promising approach as it relies on only locally available data, and re-uses calculations already made boosting the speed of computation.

Focusing exclusively on the ME assignment, we evaluate here a model which relies on an assumption that neighboring nodes of a node have pair-wise disjoint supporting arguments. The initial motivation for this kind of modeling is the 2-term order limit example of the previous section.

The formulation for this model is presented below. It amounts to combining the dsp values of neighboring nodes with the related assumption probability of the node, using a noisy-or gate. We will examine when this formulation is a good approximation later in this section.

$$d\hat{s}p_i = 1-(1-p(a_i))\prod_{j\in P_i}(1-p(l_{ji})dsp_j) = p(a_i)\,\hat{\vee}\bigvee_{j\in P_i} p(l_{ji})dsp_j \tag{5.3}$$

where $d\hat{s}p$ represents an approximation to the *dsp* value, P_i is the set of parents for node *i*, l_{ji} denotes the link assumption linking from *j* to *i*. Note that for this to be useful we have to know the dsp values for the neighboring nodes in advance, we will deal with this problem in section 5.6 when we introduce the ESP algorithms.

Let us now examine Eq.(5.3). For the ETRI setting, it follows from Theorem 3.1 that for *any* node:

$$SP_{ETRI}(v_i,\xi) = a_i \vee \bigvee_{j\in P_i}\left[l_{ji}\wedge SP(v_j,\xi)\right] \tag{5.4}$$

We exclusively refer to supporting arguments SP_{ETRI} in the context of ETRI in the following, so we drop the subscript ETRI unless we state explicitly otherwise.

If we know a priori that all supporting arguments for all parents v_j of document v_i $SP(v_j,\xi)$ are pair-wise disjoint then we can use the Sylvester-Poincare development as in Eq.(2.27):

$$dsp\ (v_i,\xi) \quad = \quad dsp\left(a_i \vee \bigvee_{j\in P_i}\left[l_{vi}\wedge SP(v_j,\xi)\right]\right)$$

$$= \quad 1-(1-p(a_i))\prod_{j\in P_i}\left(1-p(l_{ji})dsp(v_j,\xi)\right)$$

$$= \quad p(a_i)\,\hat{\vee}\,\hat{\bigvee}_{j\in P_i} p(l_{ji})dsp_j \tag{5.5}$$

This is the exact value counter-part of Eq.(5.3). This shows us the condition it should always yield exact values. We use the term ***total independence of (supporting) arguments*** to refer to this situation. As such when we assume it to be true, it is the ***total independence assumption.***

Total independence assumption is equivalent to assuming that all the paths leading to a document from its ancestors are non-overlapping. This kind of graph is actually a *tree*. Obviously this is not the reality most of the time, as say the web does not have a tree structure!

Yet still, intuitively we would expect this approximation to yield relatively better results when links amongst nodes in the underlying graph are sparse and fairly uniformly distributed.

5.4. The Common Conjunction Model for Local Approximation of dsp Values

As we intend mainly to deal with small-world networks (e.g. citation networks, web), the total independence assumption is not a good representation of the underlying structure. We have shown in section 2.2 (Watts and Strogatz, 1998)(Newman, 2003) that social networks exhibit a property known as "clustering" which basically states that:

"If two vertices in a network are connected, then a third vertex connected to one of the first two, is more likely to be connected to the other as well in a social network compared to a random network."

We use this phenomenon to basically reason that, supporting arguments for a node are more likely to be related to some extent than be completely disjoint.

To accommodate for this we re-formulate the approximation in Eq.(5.3) as:

$$\begin{aligned} \hat{dsp}_i &= 1-(1-p(a_i))d_c(v_i)\prod_{j\in P_i}(1-p(l_{ji})dsp_j) \\ &= p(a_i)\hat{\vee}\left[d_c(v_i)\hat{\bigvee}_{j\in P_i}p(l_{ji})dsp_j\right] \end{aligned} \tag{5.6}$$

where $\hat{dsp}$ is an approximation to the *dsp* value, P_i parents of i, $d_c(v_i)\rightarrow\mathbb{R}$ is a function (described below) where $0\le d_c(v_i)\le 1$.

Here we introduced the function of ***damping for conjunction*** d_c or shortly ***damping function***, which represents a hypothetical amount of "common conjunction" incident on a node.

As shown in Figure 5.1 this formulation is equivalent assuming that all supporting arguments share a common assumption and be pair-wise independent otherwise.

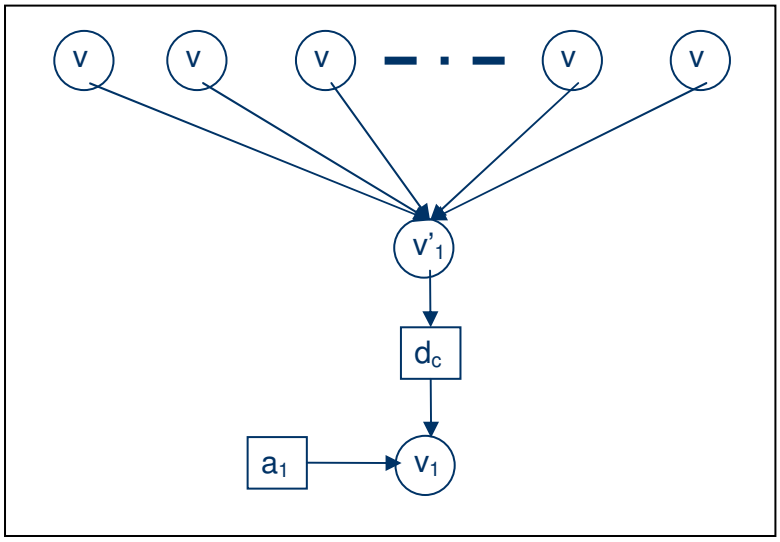

Figure 5.1. The common conjunction model

Our numerical investigations on CiteSeer scientific citation networks have shown rather high damping values close to 1. This was because link assumption probabilities we have used were rather low (e.g. constant 0.05 to 0.3), so the discounting caused by the pair-wise conjunction are limited. See section 6.5 for a full discussion of this.

An attempt to mathematically relate the small-world model parameters (especially clustering coefficients) with the damping function may be an exciting prospect, but it has been left out of the scope of this work.

5.5. The ETRI Support Propagation (ESP) Algorithms for PAS-ETRI

In this section we introduce a family of algorithms which build on the common conjunction model. The basic idea is to use dsp estimates of the neighboring nodes iteratively in a convergent scheme to calculate gradually better estimates for all the nodes step by step.

An initial problem to address is the positive feedback created by closely linked nodes – equivalents of “rank sinks” in PageRank calculations. Especially for the case of cross-

linking (e.g. when a document both cites and is cited by another document), the problem gets worse.

Firstly we introduce an iterative algorithm which has no "feedback protection". We will be mainly focusing on this algorithm because the experimental network we have (CiteSeer citation network) does not have excessive cross-linking. One may think that cross-linking should not exist in a citation network but it does, for example when two papers from the same authors may cite each other if they appear in the same journal or conference.

Then we will introduce a second algorithm which is based on message passing, and provides first-order feedback prevention (i.e. prevents feedback from immediate neighbors). This algorithm is reminiscent of Pearl's Belief Propagation (BP) algorithm (Pearl, 1988) for the ETRI framework which we discuss in section 5.7.

One can formulize higher order ESP algorithms which prevent higher order feedbacks (i.e. feedback from neighbors which are further separated by two or more links). Their structure and usefulness are left as future work and are not going to be addressed in this work.

The ESP algorithms are usable on any ETRI based mode framework, and the definitions presented here will be valid for any ETRI model. However, the evaluation for the general case of ESP is out of the scope of this work. For evaluation purposes, we will mainly be focusing on ETRI-DIM and particularly ArgRank approximations.

5.6. 0^{th} Order ESP: The Iterative Algorithm

In this section we introduce the 0^{th} order ESP algorithm, ESP-0, which is a simple iterative algorithm. We will examine its properties and usefulness by using some theorems and propositions. This is an algorithm which works much better when there is no cross-linking between nodes in an ETRI graph.

Given an initial estimation of dsp values $d\hat{s}p_i$ at any step, the algorithm iterates on this equation:

$$d\hat{s}p_i^* = 1-(1-p(a_i))d_c(v_i)\prod_{j\in P_i}(1-p(l_{ji})d\hat{s}p_j)$$

$$= p(a_i)\hat{\vee}\left[d_c(v_i)\hat{\bigvee}_{j\in P_i}p(l_{ji})d\hat{s}p_j\right] \quad (5.7)$$

where $d\hat{s}p_i^*$ represents the best-estimation for dsp values (to be used next step) given the current graph.

The pseudo-code then is as follows:

Function: *ETRI Support Propagation-0*

Input: $V = \{v_1, v_2, ..., v_n\}$, $E = \{e_1, e_2, ..., e_m\}$,

$\Pi = \{p(a_1), p(a_2), ..., p(a_n), p(l_{ij}), p(l_{kl}), ..., p(l_{mn})\}$, $d_c(v_i)$, δ

$\hat{dsp} = \{0, 0, ..., 0\}$ /* n elements */

$\hat{dsp}^* = \{0, 0, ..., 0\}$ /* n elements */

$s = 1$

do /* the iteration loop */

For each v_i **in** V do

$dsp_{tia} = 0$ /* dsp for total independence assumption */

For each v_j **in** *Parent*(v_i) **do**

$$dsp_{tia} = 1 - (1 - dsp_{tia})\left(1 - p(l_{ji})\hat{dsp}_j\right)$$

next

$$\hat{dsp}_i^* = 1 - \left(1 - p(a_i)\right)\left(1 - d_c(v_i) \cdot dsp_{tia}\right)$$

next

if *difference*($\hat{dsp}^*, \hat{dsp}$) $< \delta$ **then break**

$\hat{dsp} = \hat{dsp}^*$

$s = s + 1$

loop

Output: $\hat{dsp}^*$

Here we assume the availability of natural graph functionalities for locating parents and related links, and *difference*($\hat{dsp}^*, \hat{dsp}$) is any preferred convergence detection function (e.g. L1 norm on difference) for which δ assumes relevant meaning.

For assessing the capabilities of the algorithm, we present the following:

Proposition 4.1

Given a PAS-ETRI system $ETRI(G, PAS_{ETRI}, R)$ *and a constant function of damping for conjunction* $d_c(v_i) = 1$, *then the* ETRI Support Propagation-0 *function outputs exact results (after finite iterations) for* dsp_i *values if the underlying graph G is a tree.*

A proof of this proposition is going to be outlined here. We know that a tree necessarily has a node with no incoming edges, which we can refer to as the top node. So, we know that after the first iteration of the algorithm, we have for the top node:

$$d\hat{s}p_i^* = p(a_i) \tag{5.8}$$

which is the correct value. For the second iteration, the children of the top node will receive the dsp values from their parent, that is:

$$d\hat{s}p_i^* = p(a_i) \hat{\vee} p(l_{ji}) d\hat{s}p_j \tag{5.9}$$

This will also yield exactly correct values for those nodes because in a directed tree graph there is only one incoming path to a leaf from the top node. Also note that the top node will remain unchanged.

Our technique for the second iteration would actually be valid at any step in the algorithm. Thus, we could create a proof which would use induction to show that after some finite iterations, all the nodes in the tree would have a constant value which would be the correct dsp values.

Theorem 5.1

When run on a given PAS-ETRI system $ETRI(G, PAS_{ETRI}, R)$ *the* ETRI Support Propagation-0 *function produces nondecreasing intermediate* $d\hat{s}p$ *value assignments*

compared to the previous iteration, at the end of each iteration loop. Stated mathematically:

$$\forall v_i \in V : d\hat{s}p_i(s+1) \geq d\hat{s}p_i(s) \text{ where } s \text{ is the count of the iteration.} \quad (5.10)$$

Theorem 5.2

Given a PAS-ETRI system $ETRI(G, PAS_{ETRI}, R)$, *results* $d\hat{s}p^*$ *output from the* ETRI Support Propagation-0 *function run on ETRI, and the function of damping for conjunction* $d_c(v_i) \to \mathbb{R}$ *such that the following inequality holds:*

$$\forall v_i \in V : dsp_i \geq 1 - (1 - p(a_i)) d_c(v_i) \prod_{j \in P_i} (1 - p(l_{ji}) dsp_j) \quad (5.11)$$

This implies that the following inequality holds:

$$\forall v_i \in V : d\hat{s}p_i^* \leq dsp_i \quad (5.12)$$

Theorem 5.3

The ETRI Support Propagation-0 *algorithm terminates after a finite number of iterations, when vector difference is used as the difference function with a constant valued* δ *vector representing the desired termination values. More specifically, we use:*

$$\delta_i = e_0 \quad (5.13)$$

where i is the index of the vertex to be examined, e_0 *is the desired termination value where* $0 < e_0 \leq 1$ *and,*

$$difference_i(s) = d\hat{s}p_i(s) - d\hat{s}p_i(s-1) \tag{5.14}$$

where s is the count of iteration, and $s \geq 1$. *Then we state:*

$$difference_i(s) \leq \delta_i \tag{5.15}$$

for some $s > s_0$.

We are not limited to vector difference for convergence. For example it can similarly be shown that, the algorithm converges using L1 norm.

Thus, we have asserted that the *ETRI Support Propagation-0* algorithm has the following properties:

- Yields exact results for trees using $d_c = 1$
- The final results output by the algorithm are bound from above by the correct dsp values given a corresponding d_c.
- Each iteration may only produce better estimates of the real dsp values (for any d_c), or remain constant.
- The algorithm terminates after a finite number of iterations given corresponding termination conditions (e.g. vector difference or L1 norm using a constant termination value)

These properties imply that, given a d_c for which the inequality (5.11) holds, the *ETRI Support Propagation-0* algorithm necessarily converges to a set of values bound from above by the true dsp values, regardless of the underlying graph structure. A trivial case for this assertion is where $d_c = 0$. It can be shown that, this indeed is the case for some settings. At the other extreme is the exact solution. So at this point, intuitively we can expect that the higher the d_c value (that obeys the inequality (5.11)) the better the approximations output should be.

These theorems make no claim regarding the accuracy of the results produced. That is an issue we will address experimentally in chapter 6. Our experiments have shown us that, for many PAS-ETRI setups, ESP converges fairly quickly to a reasonable approximation given a good d_c function. We have used a constant damping function in our setups, although it is conceivable that one could come up with some heuristics to relax the constant value assumption to yield better approximations.

What we have shown here is a worst-case situation where all the *dsp* estimates obey an upper-bound set by the real *dsp* values. However, in an actual implementation relaxing this strict pre-requisite may produce better approximation results which do not necessarily observe the real dsp values as upper-bounds. As we will detail in chapter 6, we actually choose to use the average damping value (not the minimum one as the theorem suggests), which produces fairly good approximations. A theoretical assessment of the trade-off in using higher d_c values to get better approximations (against what the theorems suggest) is a topic we leave to be addressed as a future work, we will deal with it using various experimental results.

5.7. 1st Order ESP: The Message-Passing Algorithm

In this algorithm, within a step each node passes a message $\sigma^*(v)$ to all its children nodes containing its best estimation for its dsp value, and receives such messages from its parents. For the next step, a new estimate for its dsp, based on the recent messages is calculated using Eq.(5.16). The algorithm goes on until the values converge within a desired level.

Thus at a given message passing interval, using Eq.(5.5) we see that for each node the following equation is evaluated:

$$d\hat{s}p_i^* = 1-\big(1-p(a_i)\big)d_c(v_i)\prod_{j\in P_i}(1-p(l_{ji})\sigma^*_{v_i}(v_j)) \tag{5.16}$$

where $\sigma^{*}_{v_i}(v_j) = d\hat{s}p_j(v_i)$ is the message sent from the parent node v_j to the child node v_i which contains the dsp value for node j excluding the effect from node i. For the next interval, $d\hat{s}p^{*}_i$ values will be used for the messages, and so on.

This far, the reader may have noticed a similarity between the algorithm being suggested and the use of Pearl's belief propagation (BP) algorithm (Pearl, 1988) on loopy networks. In this algorithm, belief is propagated between nodes in a Bayesian network, but convergence is not guaranteed. The similarity is more imminent with the two-color model in (Broder *et al.*, 2000b). Also, it is possible to observe a graphical similarity between factor graphs and ETRI graphs, especially considering this approximation scheme proposed, and also between Support Propagation algorithm and Sum-Product Algorithm for factor graphs as in (Kschischang *et al.*, 2001). This is related to the fact that the corresponding algorithms on factor graphs and Bayesian networks are mathematically equivalent, and the graphs are mutually convertible (Yedidia *et al.*, 2003). Yet it is not possible to apply these algorithms in our setting without modification because, informally stating, for the PAS-ETRI model no Markov-Blanket localizing the "reasons" necessarily exists. So we perceive the use of ESP algorithms on ETRI networks on a Dempster Shafer theory based context (i.e. PAS), to be in a similar spirit to use of BP on loopy Bayesian networks in a Bayesian context.

In this work we chose not to focus on ESP-1 in favor of ESP-0 which has the advantage of being easier to implement and apply to bigger collections. A treatment of ESP-1 thus remains as a future work. Additionally, ESP's of higher order, effectively mixing an actual PAS implementation calculation for the micro structure (lower orders) within an ESP framework managing the macro structure are conceivable. This prospect is however, out of the scope of our current treatment of the subject as well.

5.8. Applying ESP-0 for Approximating ArgRank: ERank-0

Here we present a straight-forward application of ESP-0 for approximating ArgRank values. We use the term ERank-0 in short for applying ESP-0 algorithm to find ArgRank values on an ETRI with the ME assignment.

The main open question regarding the application is the choice of the damping function. We specify a constant damping function d_c:

$$d_c(v_i) = d_0 \tag{5.17}$$

We propose that an actual PAS-ETRI implementation should initially be run on a training set. For example, the training set could be randomly selected nodes from the actual network. Then using Eq.(5.5) the "correct" d_c can be obtained for the training set. Based on these results, a selection would then be made for the value of d_0.

Although we have proved the upper-boundedness only for the case where $d_0 \le d_c(v_i)$ for some d_c representing correct results, it is conceivable that setting d_0 to a higher value minimizing the errors for dsp values in the training set could give better approximations.

Thus using Eq.(5.7) for ESP-0, we can formulate the following iterative evaluation:

$$\begin{aligned} \hat{dsp}_i^* &= 1-(1-p(a_i))d_0\prod_{j\in P_i}(1-p(l_{ji})\hat{dsp}_j) \\ &= p(a_i)\,\hat{\vee}\left[d_0\,\hat{\bigvee}_{j\in P_i}p(l_{ji})\hat{dsp}_j\right] \end{aligned} \tag{5.18}$$

The results can then be obtained using the ESP-0 algorithm.

6. ANALYSIS OF EXPERIMENTAL RESULTS

6.1. Overview of Results

We have tried to assess the utility of our introduced algorithms, examine their results and thus gain a general understanding on their modes of working. We have made comparative analysis, and in the process we hope to have established a better understanding of both the previously known algorithms (citation count/in-degree and PageRank) and our newly introduced algorithms (ArgRank and ERank-0).

The first problem to tackle is the selection of link assumption probabilities. We evaluate two different approaches. Firstly, we use a constant value for all the links, then we use a value inversely proportional to the out-degree of a node for the links going out of that node.

Then the damping values are calculated. We have calculated actual ArgRank values to orders between 3^{rd} and 5^{th}, and used them to approximate damping values for a sample set. We chose to use the average damping values of the sample sets as damping values for the ERank-0 algorithms.

Our choices for link assumptions and the damping values, have given us a total of three different parameter settings to evaluate for the ERank-0's; (a) (b) and (c). With the addition of CitationCount and PageRank, we have run a total of 5 different algorithms on our data. There is an additional (c2) setting. In this we have applied the (c) setting to the pruned version of the network which we used for applying PageRank. (see section 5.9 for further explanation)

Our first effort after the ERank-0 runs, was to evaluate the accuracy of obtained results. The very reason we have proposed using an approximation has, not surprisingly, caused for us the problem for assessing the quality of our obtained results. It can become

exceedingly difficult to compute realistic estimates of ArgRank values when terms with high orders are necessary. As we will demonstrate in this chapter, we have evidence that higher order terms (e.g. higher than 5) may indeed become the dominant factors on determining the ArgRank values for networks where the global influences dominate.

To attain an understanding on the character of the network we deal with, we have examined the small-world network model characteristics such as, in-degree distributions (the scale-free property), average distances and diameter (the small-world property). We confirm with reasonable confidence the previous findings, we also find that similar properties are exhibited by some of our introduced algorithms.

We introduce the transition of dominance between global vs. local influences as an emergent characteristic on which to assess the results of different algorithm settings. In this context, the CitationCount algorithm (in-degree), becomes the extreme end for incorporating local-only data, and using two ERank-0 settings (a) and (b) we try to explore the effect of more global influences.

In similar sprit as CitationCount is similar to ERank-0(a) and ERank-0(b), we use ERank-0(c) as an "evidence based" analog for PageRank.

We present comparative analysis between the algorithm settings, in different forms such as: various scatter plots, average position distances (introduced in this chapter), and correlation coefficients.

A detailed assessment of the convergence characteristics was out of the scope of this work. In our experiments we have observed that the convergence pattern is highly related to link assumption probability assignments. The higher valued and globally dominated settings take longer to converge to similar levels relatively. For example the (a) setting converges in 10 iterations, whilst the others can take more than 50. We have assessed the level of convergence by comparing the results with a previous iteration. An interesting observation was that for setting (b) the results initially diverge (the difference between two

consecutive results increase) for a while and then attain a converging trend (monotonically decrease).

We use the top-ranking documents as a demonstration of our results, which we hope can be useful for getting an idea of the promoted documents by the different algorithms given the subjectivity of the topic we deal with.

One of the essential uses of an algorithm assessing information value, is assisting the information retrieval process. As we argue and find suggesting evidence later on, the global picture may not necessarily give a good understanding for the experience of the information searching agent. So, we further our analysis in this direction, and give similar analysis on a per-query basis.

6.2. Overview of Data: CiteSeer Citation Network

We have run our algorithms on data based on the CiteSeer (CIT) online paper collection. CiteSeer is an open online database, and makes available scientific literature mostly on computer and information sciences.

We have used the metadata provided by the archive as part of the Open Archive Initiative (OAI) . It is a snapshot from 03-2005 based on an extended version of the Dublin core standard including citation information along with some other useful additional meta-data fields.

While being a fairly large set of documents, we have observed our collection had some short-comings some of which are certainly shared with any limited collection one may examine. Upon examining the result sets for some queries we have seen that, it is possible some of the most influential papers along with others on a topic may not be present in the collection. Also as being a best effort project, essentially indexing and collecting literature freely available on the net, the citation information is not always complete. This is possibly due to a failure in the automation process which extracts this citation data from the papers' text.

Our data downloaded from the CiteSeer web site included information on only those papers that were actually hosted with full content within the collection. It was not possible to obtain the much more useful data which includes the "context" data which includes the references – but not the content – of a document.

Unavoidably these must have degraded the quality of the rankings as we solely rely on citation data on documents. Nevertheless, our impression has been that the collection does contain a subset of the influential papers and reasonably accurate citation data, so this enables us into asserting with some reservation that our results are representative.

The ETRI network is constructed such that, whenever a document "a" references a document "b" there is a link from node "a" to "b". This actually is one of the numerous ways to construct such a network. However, given the amount of effort necessary to run our analyses on a network, we have opted to concentrating our efforts on this single structure for the scope of the work.

Some of the characteristic values of the network are listed in Table 6.1. The distribution for the CitationCount algorithm (in-degree distribution) is presented along with the other algorithms.

Table 6.1. CiteSeer (forward) citation network properties

Name of the value	Value
num. of vertices	299 772
num. of (directed) edges	1 255 566
average path length	23
diameter	74
power-law exponent	3.01

6.3. The Experimental Setup

To facilitate a comprehensive analysis of the collection, we have effectively built a document search engine. We have used a relational database system (an SQL server) for storing and indexing various data on documents.

Our setup enabled us to run queries on composite data consisting of descriptions/abstracts of documents (the first 1000 characters), authors and titles. It also enabled structured manipulation and analysis of our data, in many useful ways (e.g. sorting, pruning, ...).

We did not need a text matching based similarity measure, so it sufficed to have a basic keyword based matching in a boolean mode which was provided by the system.

We have used the open source Java Universal Network / Graph (JUNG) framework as the basis framework for implementing our algorithms in Java (JUN).

We ran into performance and memory problems while dealing with our network data, thus we had to re-write many of the core classes to suit our specific algorithm and performance needs. Despite the difficulties however, the JUNG framework has been very useful and instructive in defining a working software abstraction for dealing with networks while incorporating flexible manipulation and data holding capabilities. It also provided us with working code for importing and exporting data, and example algorithm implementations.

Also, as we detail in Appendix B, we have constructed a working PAS system capable of analyzing ETRI graphs constructed within the JUNG framework. We have used the open source JavaBDD package (JBD) for this purpose, which has been of great help by providing us with an out of the box working BDD implementation.

6.4. Choice of Link Assumption Probabilities

The calculation and choice of parameters are important issues for the application of any of the algorithms we have used.

Firstly, for ArgRank and its approximations the way of assigning link assumption probabilities need to be determined. This may not be a problem solved in a straight forward manner. As mentioned in chapter 4, the concept of information value is not a directly observable value in an objective way. That is to say, the quality of a scientific paper can not be assigned a numerical value, even after it is read thoroughly. It may not even be possible to assess the quality of it with current scientific knowledge, and it may require some time before a realistic understanding of its qualities can be well understood.

A parallel discussion can be made for the link assumption probabilities. It is not possible to accurately assess how much informative value an author attributes to a paper s/he references. However much uncertainty pertains to its assessment, it is still natural to think of citations as an evidence for the informative quality of paper (e.g. being good).

We have taken two approaches to address this problem. The first is the approach taken in (Picard, 1998) and (Picard, 2000). For this a constant value is assigned on every link in a collection. So, for the DRM, this can be thought of as our conceived evidence that a certain ratio of the papers are likely to be related out of the reference list of a paper, given the referencing paper is relevant. In (Picard, 1998), the authors develop a method for extracting a sensible value for link assumption probabilities. It is important to note that these values are specific to the sort of network they relate, and there is reason to expect that they will change between different type of networks such as the web and a citation network – maybe even within different types of citation networks.

There is a difference in the DIM compared to DRM. As we have discussed above it is not possible to objectively estimate the information value attributed to a reference. So, the methodology used for DRM in (Picard, 2000) (hyper-text retrieval model as referred to in that work) is not directly applicable to DIM in this sense. Acknowledging the ambiguity,

we chose to use two different constant values for link probabilities: 0.05 and 0.3. These may be thought of as corresponding to the assumption that, at least five per cent and 30 per cent of referenced papers are highly regarded papers (documents). Note that, as we are dealing inherently with PAS based on positive literals (positive evidence), we specify only the lower-bounds for the probabilities hence the use of "at least" 5-30 per cent. The higher-bound (the plausibility) is always 1.0. As we detail in the following sections, these two different damping values gave significantly different results.

As a reasonable method removing the ambiguity, we propose a second way of assigning link assumption probabilities. In this scheme every link gets a value inversely related to the number of outgoing links. More specifically:

$$p(l_{ij}) = 1 / N_i \tag{6.1}$$

where N_i is the number of outgoing links from a node (out-degree). Interpreting this assumption assignment in the evidence context; it corresponds to the evidence that for each document at least one of the referenced documents should be a "good" one.

One may have noticed a reminiscence of this way of assigning probabilities and the PageRank algorithm. Indeed, the results yielded by these two algorithms are indeed very similar.

A list summarizing the link assumption probability settings is given in Table 6.2.

Table 6.2. Link assumption probabilities

algorithm	link assumption probabilities
ArgRank(a)	constant value **0.05**
ArgRank(b)	constant **0.3**
ArgRank(c)	$1 / N_i$
ERank0(a)	constant value **0.05**
ERank0(b)	constant value **0.3**
ERank0(c)	$1 / N_i$
PageRank	$1 / N_i$
CitationCount	N / A

6.5. Calculation of Damping Constant

Once the choice of assigning link probabilities is made, the damping constant values are needed to run the algorithm. The calculation of a damping value requires the dsp calculation of a node to at least 3rd order, better yet to even a higher order.

We have constructed a PAS implementation, geared towards analyzing ETRI networks using Binary Decision Diagrams (BDDs) (Bryant, 1986). This system enabled us to analyze documents with a few hundred supporting arguments. The interested reader may refer to Appendix B for further information on this implementation.

As we have earlier discussed, a direct calculation of ArgRanks (dsp values) for a relatively bigger network is not feasible – if not impossible. In our example CiteSeer network, the number of supporting arguments of the 3rd or higher order for a highly cited paper may easily explode up to 100 000s, where it becomes virtually impossible to compute the dsp values with our current capabilities. Facing these difficulties, we have opted to finding high order dsp values for nodes with relatively fewer citations. We sampled approximately 200 nodes for each link probability setting. We calculated dsp

values of 3rd to 5th order supporting arguments in this group. It is an open question whether this choice should have affected our estimates of the damping values.

After a document's ArgRank of n^{th} order is calculated, the ArgRank values of its immediate neighbors for n-1st order are calculated. Then simply reversing the common conjunction model formula of ERank-0 (Eq.(5.7)), the following equation gives the damping value estimation:

$$d_c(v_i) = \frac{1 - \frac{1 - d\hat{s}p_i}{(1 - p(a_i))}}{\prod_{j \in P_i} (1 - p(l_{ji}) d\hat{s}p_j)} \tag{6.2}$$

We have chosen to use the average damping values from the sample set for use in the calculation of ERank-0 rankings. We acknowledge that, more research and justification for the methodology on computing the damping constants on sample data would be beneficial.

We have found out that the minimum damping value in all three cases is a pair of cross-linking documents, whose only citations are each other. As we have discussed in section 5.6, ERank-0 is not capable of dealing with direct positive feedback, this is a direct consequence of this short coming.

For applying the PageRank algorithm, we have used the value 0.85 as it was the most frequent damping value present in the literature. We know that this value affects the stability of the results and we are not aware of any reason why there should be a change in this value when applied to a scientific citation network as opposed to web.

To sum up, we have used four different damping values (including PageRank) to use in all our experimentations. These settings are listed on Table 6.3.

Table 6.3. Damping values for algorithm runs

Algorithm setting	Constant damping value
ERank0(a)	0.9982986
ERank0(b)	0.9737562
ERank0(c)	0.9862921
PageRank	0.85

6.6. Evaluating ERank-0 Approximation Results

We have used the ArgRank samples as a measure of comparison to ERank-0 approximations. Also included are 2nd order ArgRank values to give an idea on the effect of additional orders used.

One should note that, with ERank-0 each iteration corresponds loosely to an order of ArgRank. In ERank-0 calculations we have made, there were settings requiring at least 50 iterations. Thus, also recalling that the CiteSeer graph has an average path length of 23, even a 5th order ArgRank calculation is not necessarily a good approximation to the true ArgRank value.

In figures Figure 6.1-Figure 6.4 we display the comparisons. As can be seen in these figures ERank-0 values can be a close match to ArgRank values in some cases, while in others we may get rather higher values for ERank-0's.

ArgRank(a) values match ERank0(a) values closely in all orders (3, 4, 5) while we see that for ArgRank(b) and ArgRank(c) calculations, there is a bigger deviation for 3rd order ArgRanks.

For link assignment (b), the differences get so big that it was not possible to include 3rd order ArgRanks in a graph.

Note that, the fact that a document's ArgRank was calculated to, say, 3rd order and not 4th is simply because it was not possible to do so, due most probably to a rapid increase in the number of supporting arguments of the document. So, the deviation between ArgRank3 and ERank-0 values is mostly an expected result given this.

We observe that for the documents exhibiting bigger differences between ERank-0 and ArgRank values, their corresponding PageRank values are also higher when compared to neighboring documents sorted according to ArgRank values. We have noted from the data also that they do not have a particularly high amount of citations (e.g. 10-30). This can be taken to suggest that, these may be the documents with relatively fewer citations, which are located in "dense" areas of the network. That is to say, their referencing documents may be of higher information values (i.e. ranks), so that even with fewer citations, these documents may be gaining higher ranks.

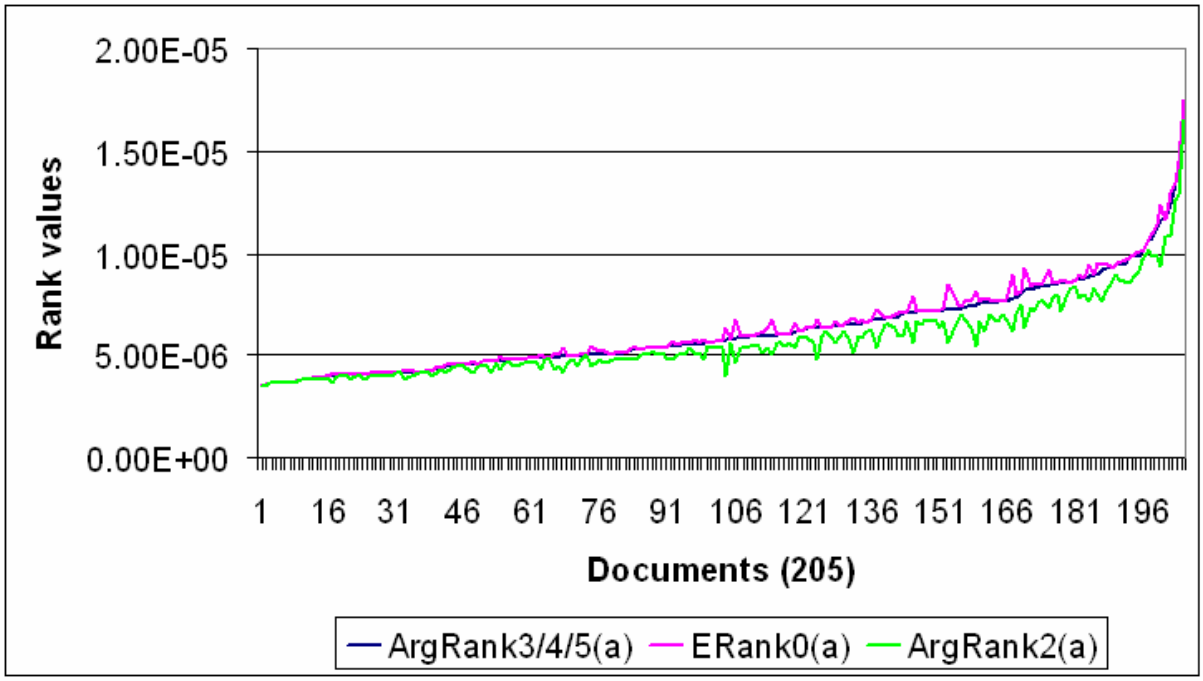

Figure 6.1. Comparison of ArgRank3/4/5 and ERank0(a)

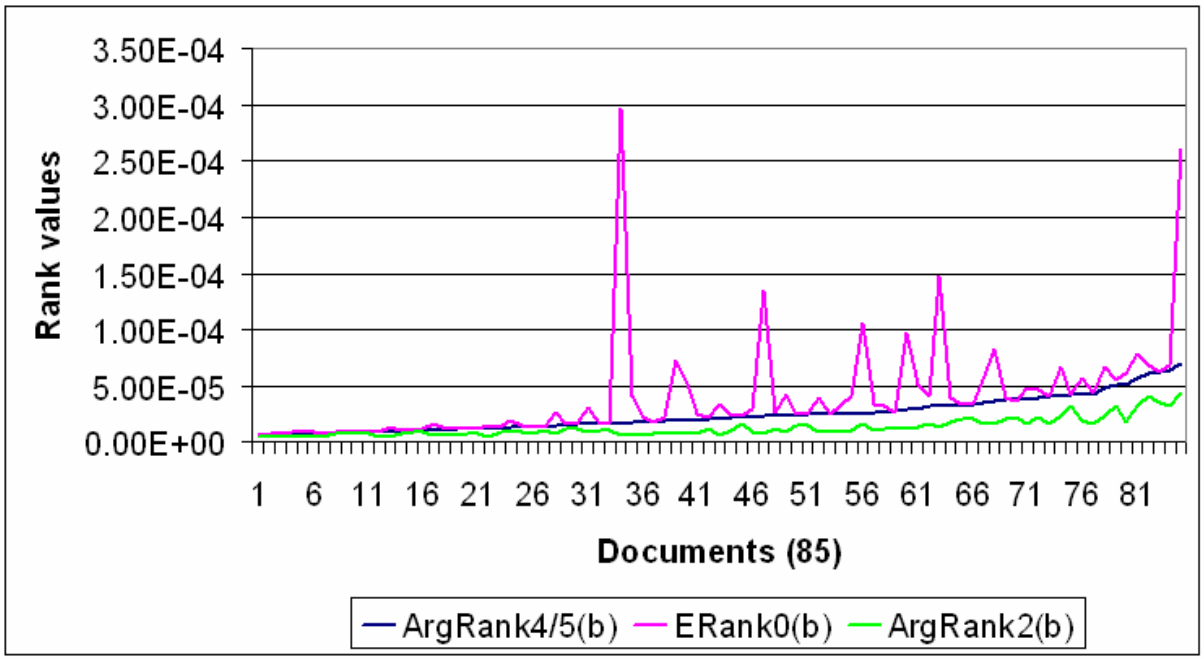

Figure 6.2. Comparison of ArgRank4/5(b) and ERank0(b)

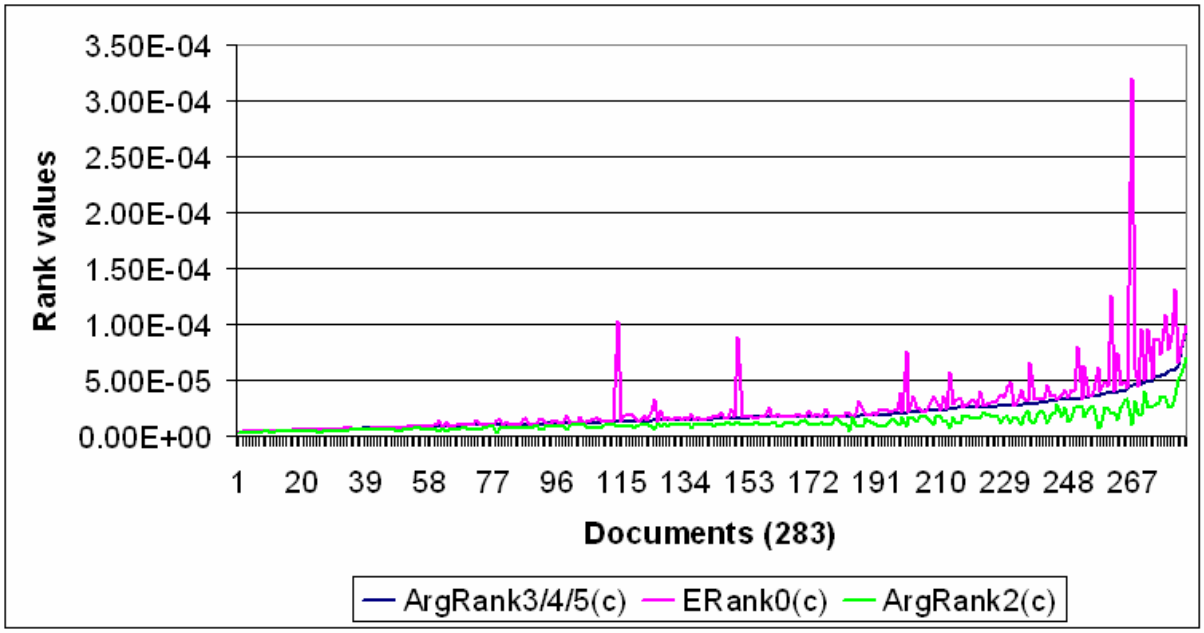

Figure 6.3. Comparison of ArgRank3/4/5(c) and ERank0(c)

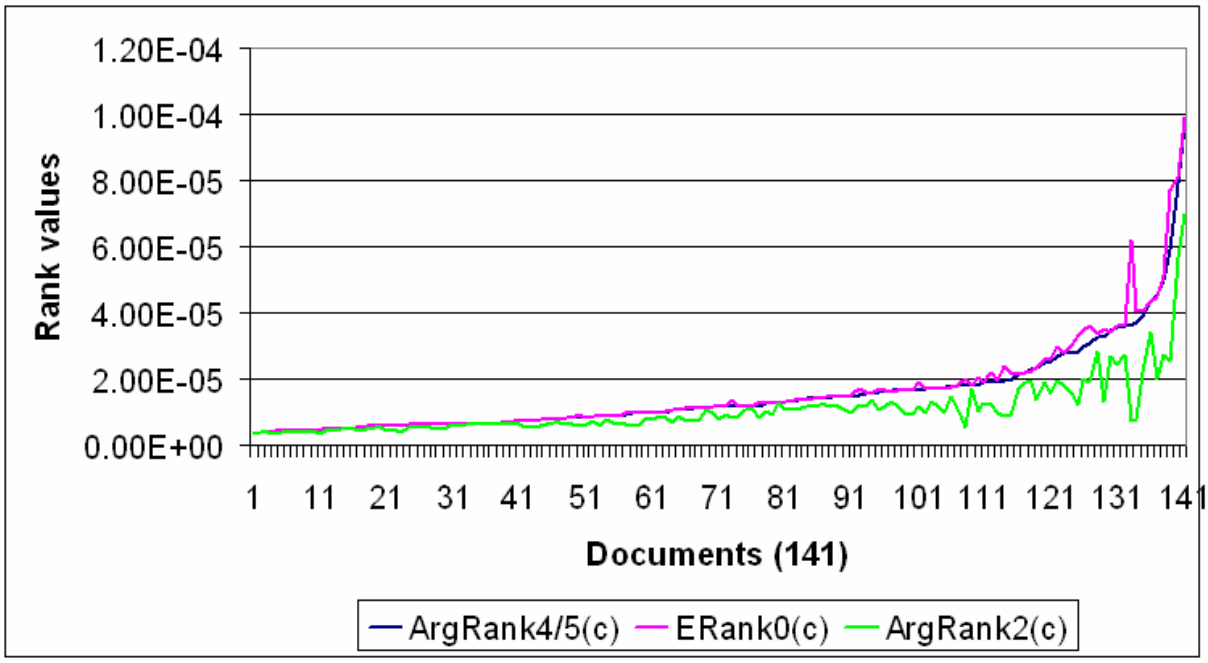

Figure 6.4. Comparison of ArgRank4/5(c) and ERank0(c)

To investigate the higher differences we observe between ArgRank(b) and ERank0(b) compared to ArgRank(a) and ERank(a) we present Figure 6.5 and Figure 6.6. Figure 6.5 is a log-log scatter plot of absolute differences between ArgRank and ERank values between settings (a) and (b), for which we find a reasonable correlation with a coefficient of 0.7130. Figure 6.6 is a semi-log plot showing corresponding log absolute differences for each document. In these figures we also see that the higher link probabilities for setting (b) occasionally result in higher than expected disturbances in differences. Obviously this is not due to noise – as these are not measurements – but

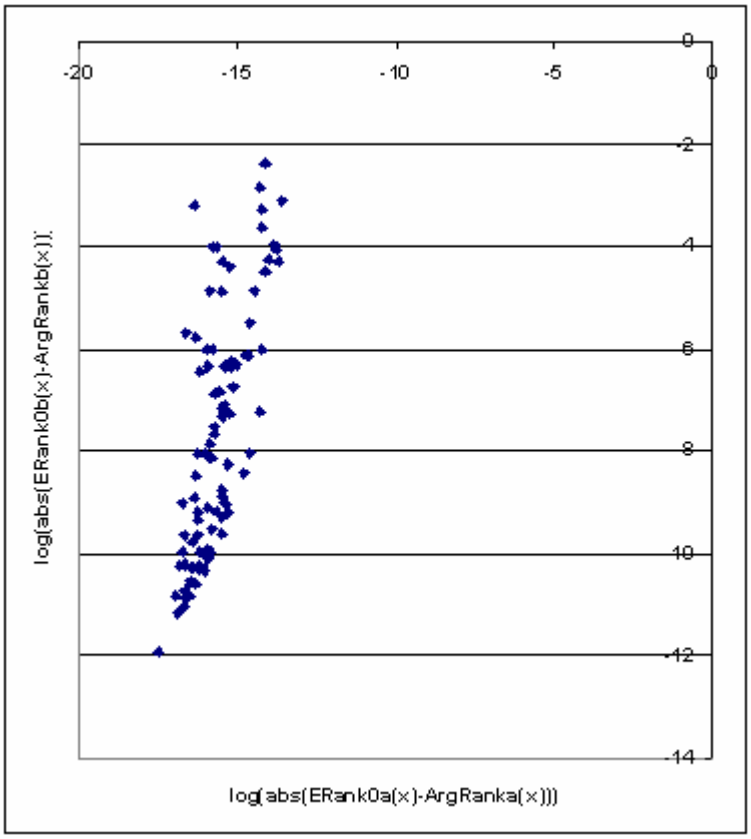

Figure 6.5. Log-log plot of differences of (a) and (b) settings

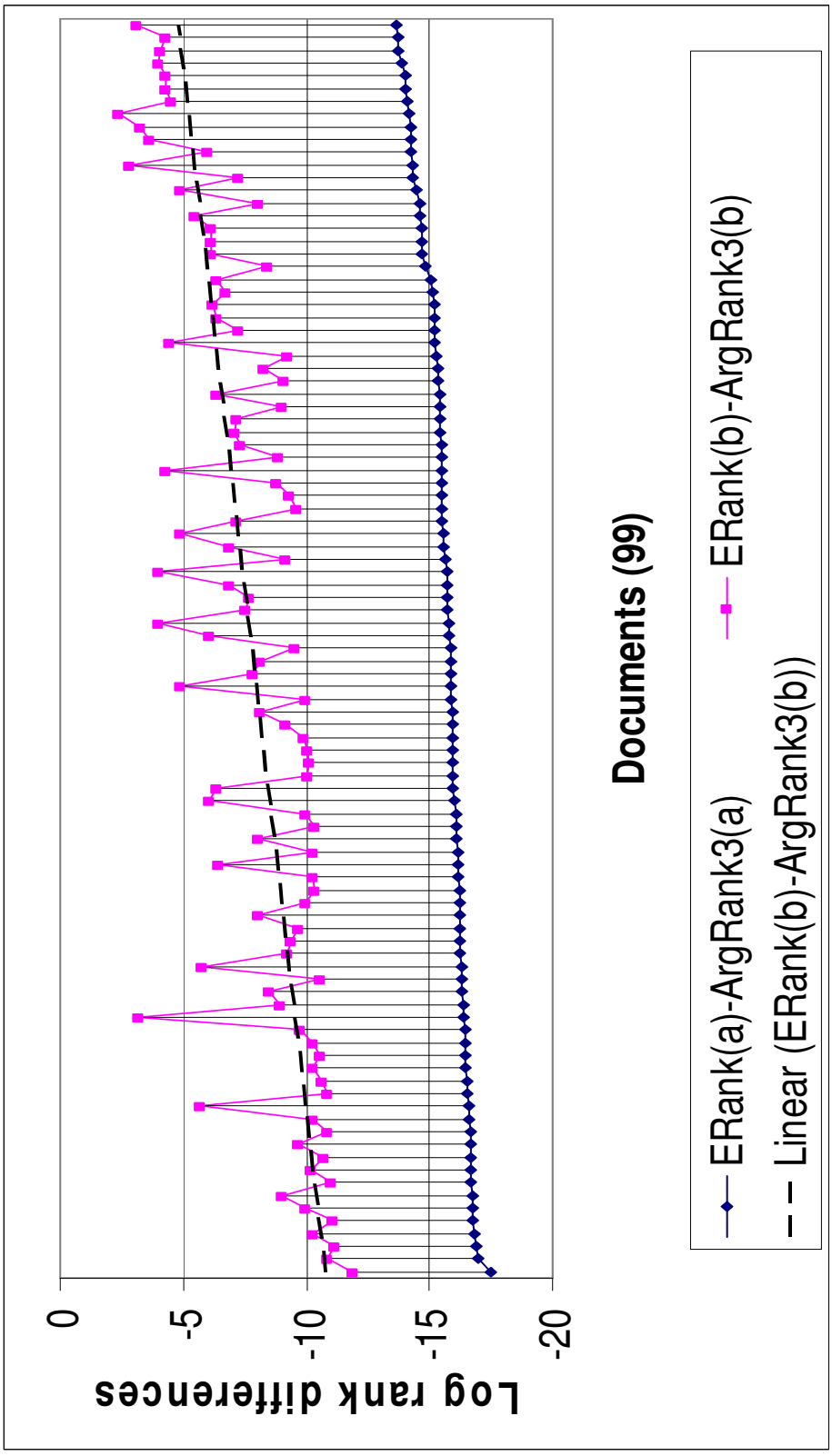

Figure 6.6. Log rank differences between (a) and (b) settings

as we will show suggestive evidence in the following sections, this is due to the domination of the macro structure (e.g. overall density of dsp values in the neighborhood of the node) of the network, over the micro structure (immediate neighbor's contribution). We recall that, it is not possible to directly establish this by calculating dsp values of higher order due to resource limitations. We employ indirect measures in the following sections to explore this issue.

6.7. Distributions of Ranks

In this section we examine the distribution of rank values produced by the various algorithms we have run. This may help to develop a better understanding on the working of these algorithms.

6.7.1 Citation Count Distribution

Examining the log-log plot of the in-degree distribution (Figure 6.7) and the corresponding Zipf plot (Figure 6.8), we see that our findings are inline with previous findings on scientific citation networks.

When the data are fit directly over the scatter plot (Figure 6.7) in the citation range between 18 – 56 (on the graph from 1.25 to 1.75) we get the exponent as -2.16. When a Zipf plot is used (Figure 6.8) we get the exponent value $\alpha \approx 3$ as detailed below. The line shown has a -3 slope, presented as a visual aid.

In the Zipf plot (Figure 6.8), we have used the first 3000 documents with citations ranging from 1404 down to 54. This best fit is the line with a slope of -0.4968, parallel to the line shown in the figure as a visual aid. It corresponds to a power-law exponent value of $\alpha = 1 + 1 / 0.4968 = 3.0129$. Fitting instead for the top 316 documents ([0,2.5] on the graph) yields a slope -0.4062, and an exponent 3.4618.

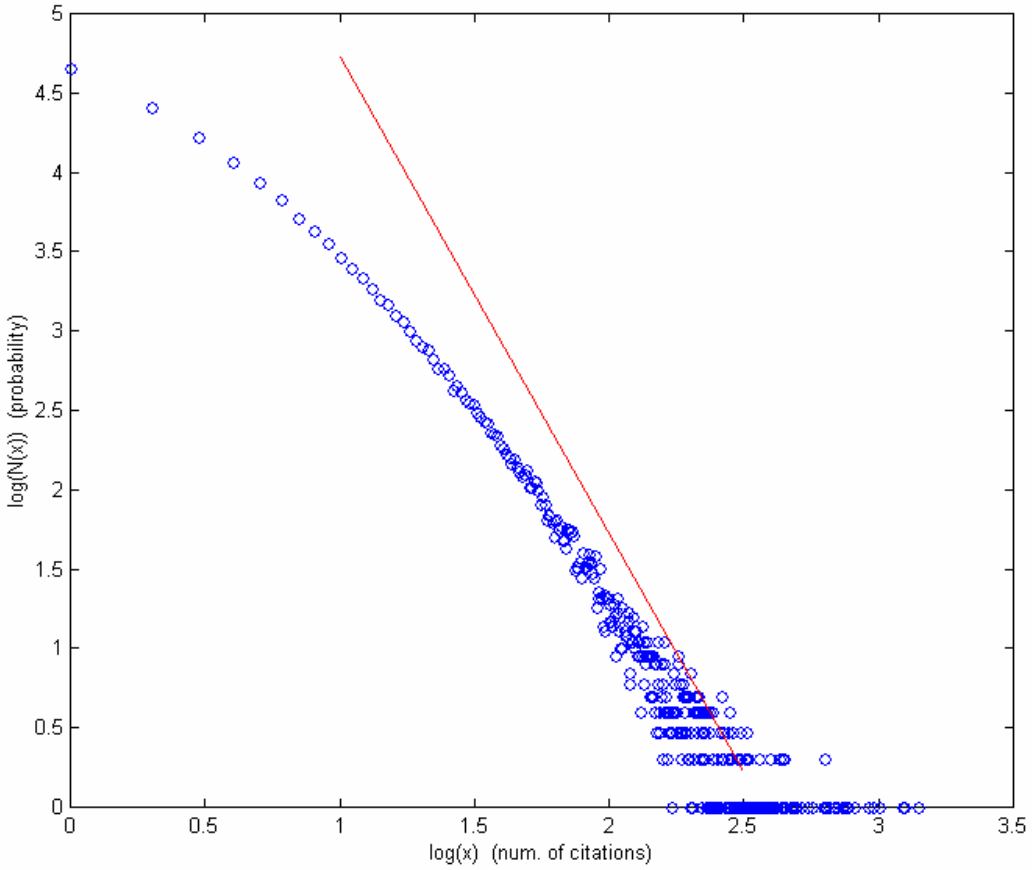

Figure 6.7. Log-log plot of citation count vs. probabilities

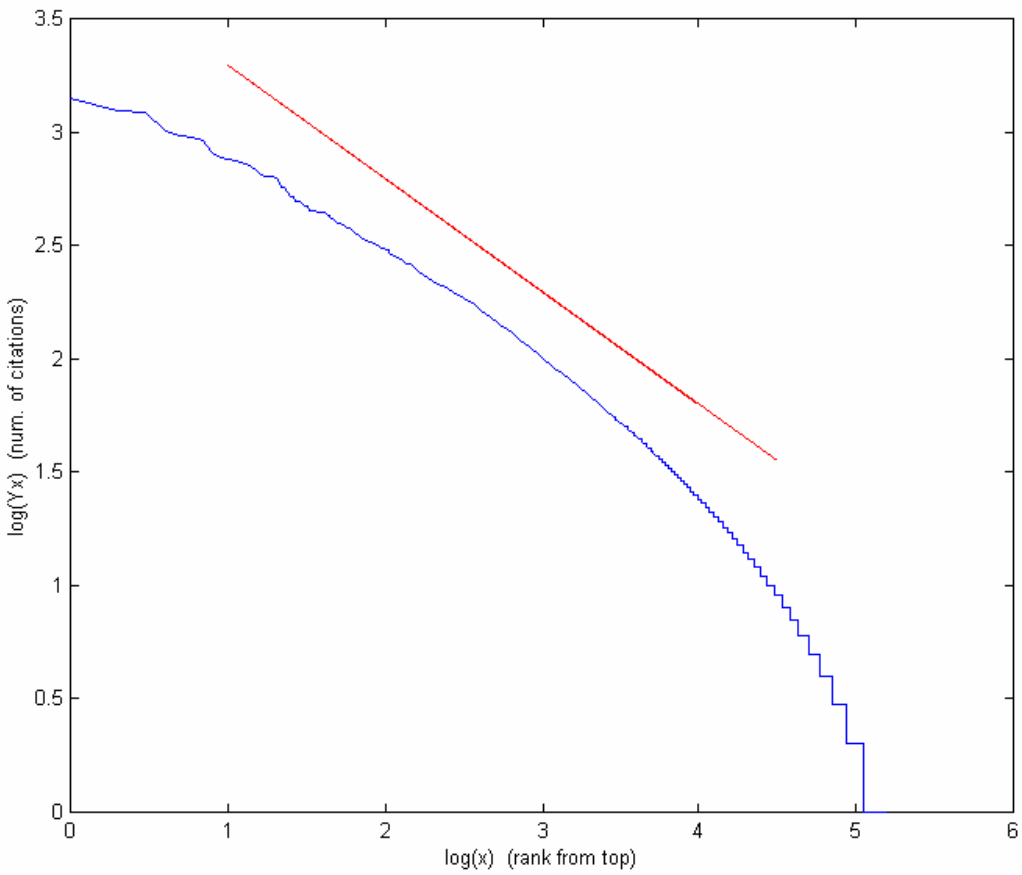

Figure 6.8. Zipf plot for citation count

6.7.2 ERank-0 Distributions

We have found that, ERank0(a) and ERank0(c) settings exhibit a power-law distribution, while ERank0(b) does not. This shows that, the different manner of assigning link assumption probabilities does not have an effect on this difference.

As can be seen in the Figure 6.9, ERank0(a) has some fluctuation in the beginning. A straightforward fit using top 1000 documents into consideration yields a line with slope -0.4185 (shown in the figure), with a corresponding power-law exponent of 3.3895. Instead if we fit between top documents range [4,158] ([0.5,2.2] in the graph) we get a slope -0.3925 and a corresponding exponent 3.5478.

We can see that ERank0(b) does not exhibit a power-law distribution in Figure 6.10.

ERank0(c) Zipf plot (Figure 6.11) has fluctuations in the beginning, so a direct line fitting is not reasonable. To by-pass the fluctuation a visual scanning of the curve we have used different ranges:

[32,316] ([1.5,2.5] on graph) gives slope -0.9050, and exponent 2.1050
[32,1000] ([1.5,3] on graph) gives slope -0.8819, and exponent 2.1339 (shown on graph)
[32,10000] ([1.5,4] on graph) gives slope -1.0072, and exponent 1.9928

We conclude that a realistic exponent value is around 2.1.

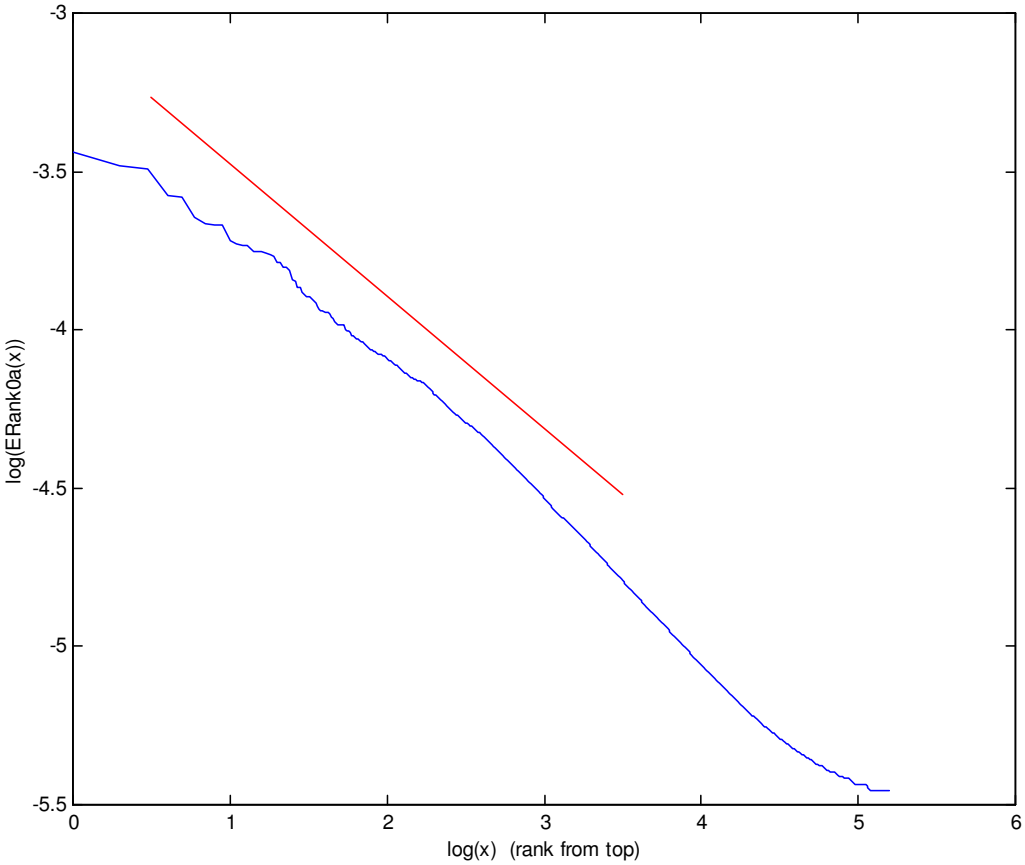

Figure 6.9. Zipf plot for ERank0(a)

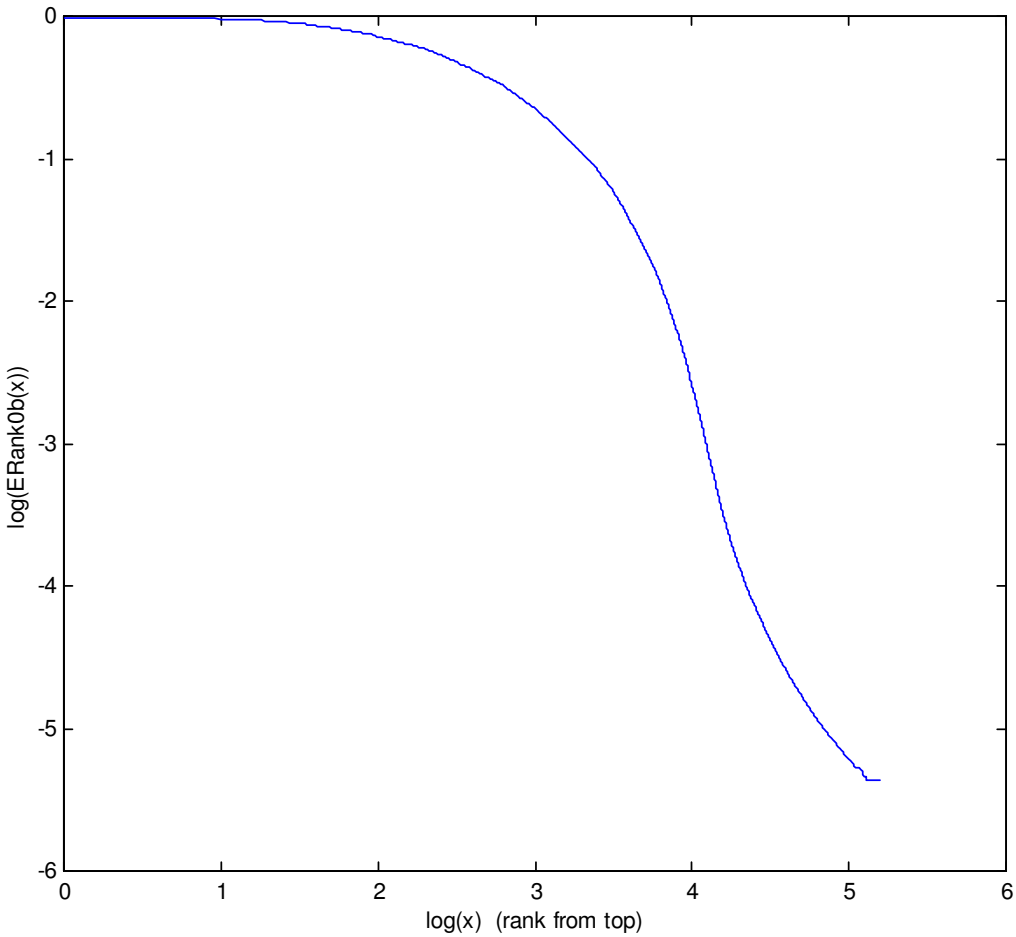

Figure 6.10. Zipf plot for ERank0(b)

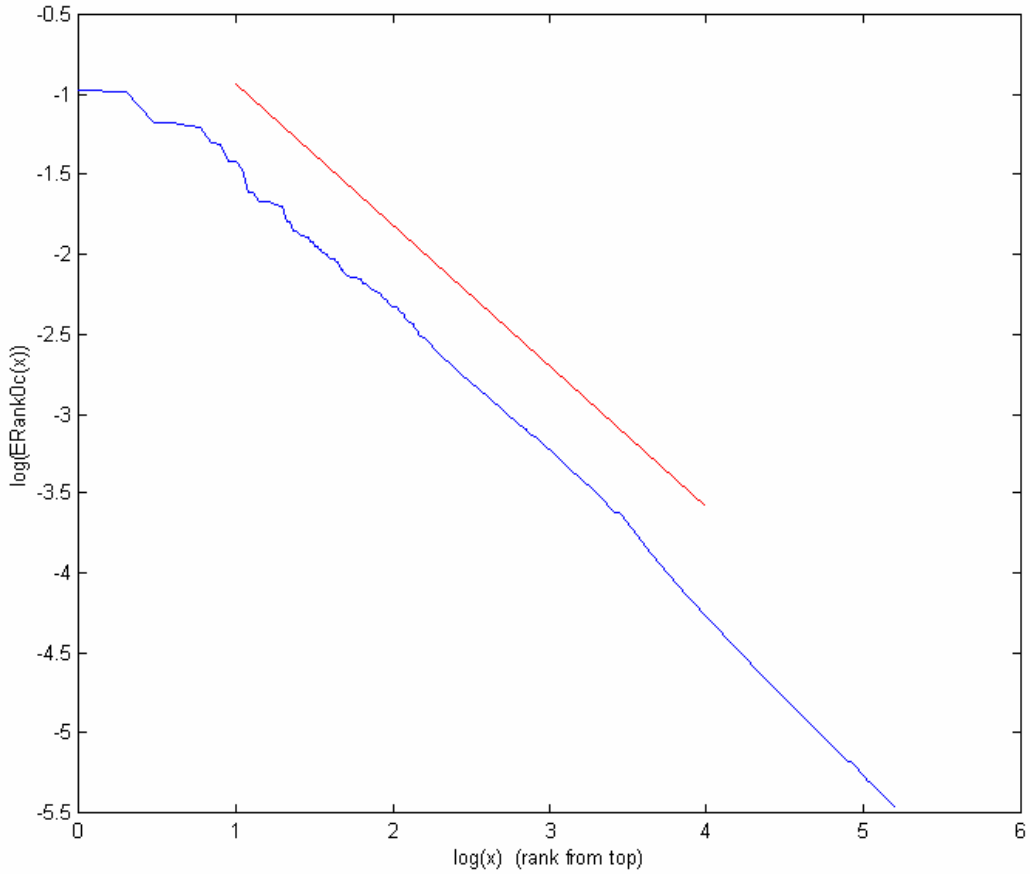

Figure 6.11. Zipf plot for ERank0(c)

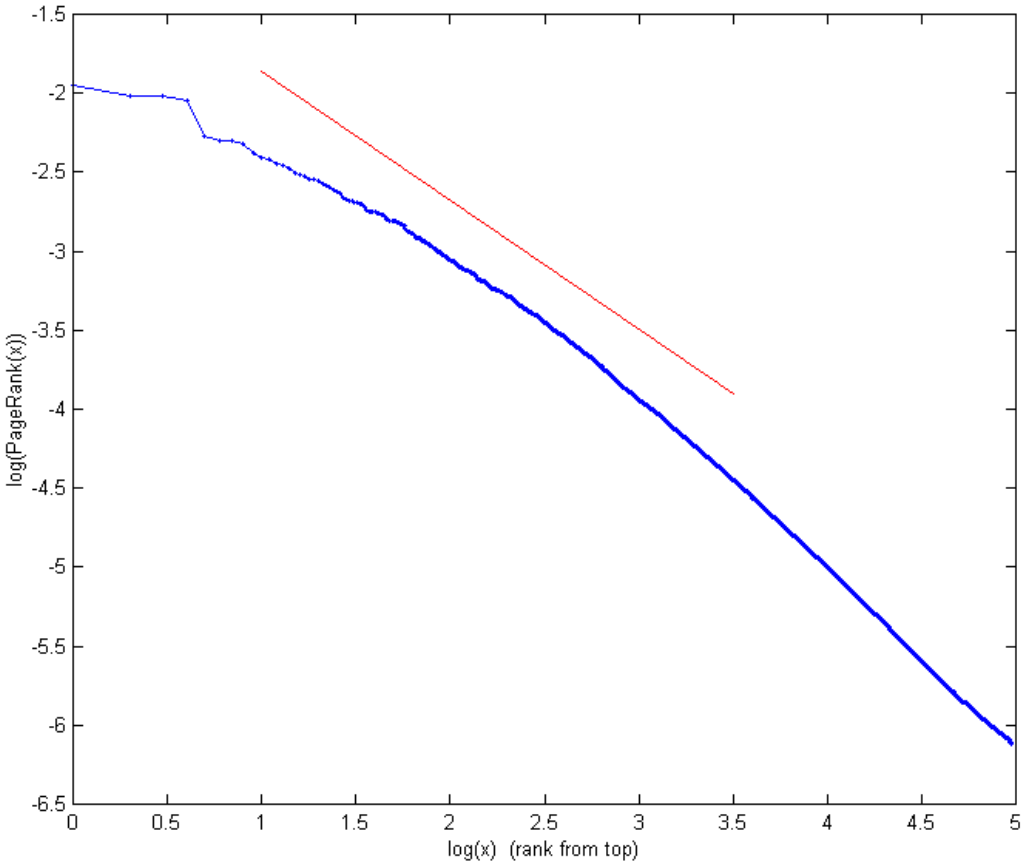

Figure 6.12. Zipf plot for PageRank

6.7.3 PageRank Distribution

As seen in Figure 6.12, the PageRank distribution appears to be fairly linear. Fitting a line for the tops 1000 documents yields a slope of -0.8155 with the corresponding exponent 2.2263 (shown on the graph). Fitting instead for the top 316 documents ([0,2.5] on graph), give a slope -0.7049 and exponent value 2.4186. A previous study on a web graph had revealed an exponent value of 2.1. This difference is expected, as these two graphs have different characteristics for in-degree distributions as demonstrated previously.

6.8. Comparison of Algorithm Results: Global vs. Local Influences

We have made a comparative study of the algorithm settings we have run on our test data. For this purpose we have employed multiple techniques. In the following sections we make a case which we build on contrasting the global vs. local influences the algorithms incorporate. As we have earlier mentioned, in this context citation count represents the local extremum whereas PageRank is covered as an established ranking algorithm incorporating global influences. We have intentionally designed our settings so that ERank0(a) is closer to CitationCount yet using some global influence, and ERank0(b) "over-emphasizes" the link evidence so that we end up having global influences dominate in the results. This is vividly demonstrated in the log-log scatter plot of ERank0(b) vs. CitationCount (Figure 6.16), where we see that a document with one citation may be ranked well above about half of the collection. Because of the way we have designed the link assumption settings in ERank0(c) in a similar spirit to PageRank, we have expected the two to exhibit some similarities as well.

It is possible to get a rough overview of our results one can consult Table 6.4 for the correlation coefficients matrix. It initially appears that, our expectations have received reasonable backing. For example ERank0(a) and CitationCount are tied with a coefficient 0.98. We try to deepen our findings using scatter plots, and later introduce a measure we call average position distance.

There is a second table for correlation coefficients (Table 6.5). This is because; we had to apply the PageRank algorithm to the pruned version of our network, in which we had to iteratively remove the 0 out-degree nodes. As we present in the background survey, this is necessary to prevent "leaking" of the ranks. This resulted in the loss of about 1/3 of the nodes. The nodes which were gone were occasionally important ones (e.g. with high citations). In contrast, all the other algorithms are applicable to the whole network, so for presenting comparisons with PageRank we use this pruned version of the network with altered results in Table 6.5. Also, minding the similarities between PageRank and ERank0(c), we have made a second run of ERank0(c), which we have called ERank0(c2) on the pruned network, to get a deeper understanding of the extent of their similarity. It can

be seen on the table that, this (c2) setting has a strong relation to PageRank with a coefficient of 0.96.

It may not suffice to depict a global picture of the results to understand the underlying structure. This is a problem we try to address in the following sections after dealing with the global picture, using query results.

Table 6.4. Correlation coefficients

	ERank0(a)	**ERank0(b)**	**ERank0(c)**	**CitationCount**
ERank0(a)	1.0000	0.5176	0.1701	**0.9770**
ERank0(b)	0.5176	1.0000	0.1509	0.4310
ERank0(c)	0.1701	0.1509	1.0000	0.1551
CitationCount	**0.9770**	0.4310	0.1551	1.0000

Table 6.5. Correlation coefficients for the pruned network

	ERank0(c2)	**PageRank**	**CitationCount**
ERank0(c2)	1.0000	**0.9079**	0.1550
PageRank	**0.9079**	1.0000	0.2848
CitationCount	0.1550	0.2848	1.0000

6.9. Comparative Plots

In this section we present scatter plots between settings in an attempt to gain an understanding of their relations.

6.9.1 CitationCount vs. ERank0(a)

In Figure 6.13 we can clearly see that some documents with fewer citations are favored over documents with more citations. Both in the high and low citation zones the

hovering effect created by the ERank0(a) setting is observable. The log-log plot is also included (Figure 6.14), it gives a better picture of lower citation count zone. For example, in the lower zone we see that a document with one citation was ranked as higher than many documents with 20 citations. Thus we clearly see how the global influences brought by ERank0(a) algorithm affect the local rankings. Yet it is fairly apparent that, the results from ERank0(a) setting increase along with increasing CitationCount (in-degree) values.

Shown lines on the graph are visual aids. On the linear plot (Figure 6.13), the line corresponds to the minimum ERank0(a) value a node receives on the graph given the citations. Similarly, on the log-log plot there is a line with slope 1.0 designating the linear relationship for the data.

Yet we note that although this plot exhibits the effect of the ERank0(a) setting well, does not provide an insight into how the internal rankings within a topic or community are affected.

From a practical point of view, this internal ordering can be more important. Because a cognitive agent would more often than not, be after reaching the most informative and relevant documents regarding a query or a topic, and not comparing the relative importance of different topics for the collections (e.g. size of the communities). This is a problem we try to address using per query analyses.

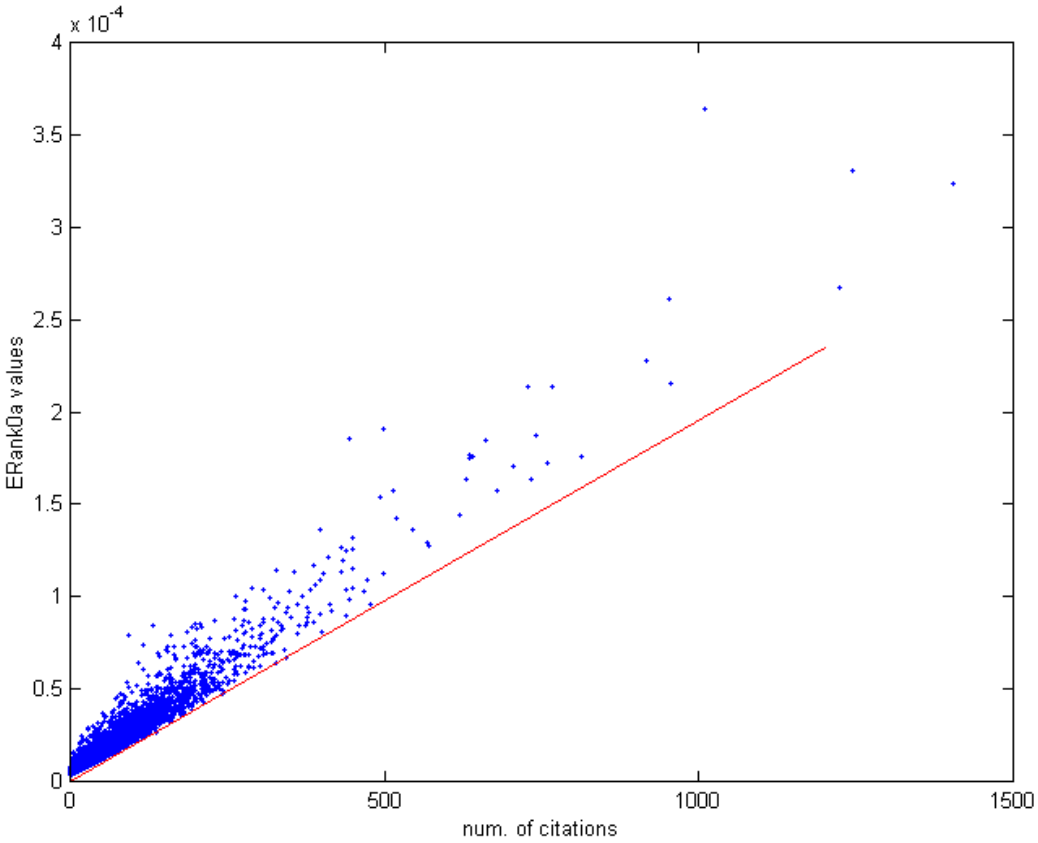

Figure 6.13. Scatter plot for citation count vs. ERank0(a)

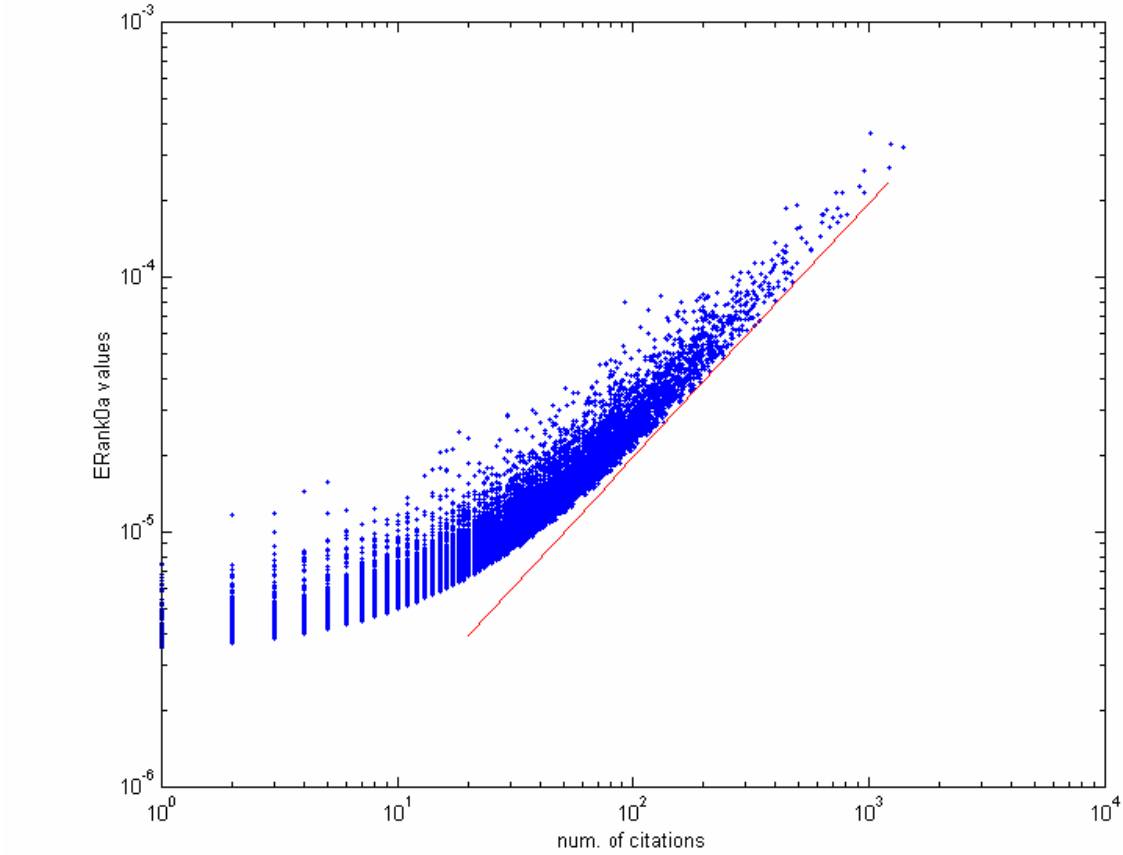

Figure 6.14. Log-log plot for citation count vs. ERank0(a)

6.9.2 CitationCount vs. ERank0(b) and ERank0(c)

We see that the linear plot of CitationCount vs. ERank0(b) (Figure 6.15) gives an apparently unsystematic relation. When we examine the log-log plot though (Figure 6.16), there appears some influence albeit small, caused by citation counts.

Comparing the ERank0(a) and ERank0(b) plots, we see more scattering in the (b) plot, which is also in line with our expectations, as the higher confidence we have in the link assumptions, the more influential community structures get. Thus, fewer citations have more influence and selectively hover some documents due to global influences.

A very similar situation is obtained by the CitationCount vs. ERank0(c) plot (Figure 6.17). The plot on the log-log plot shows the linear relation as a line with slope 1.

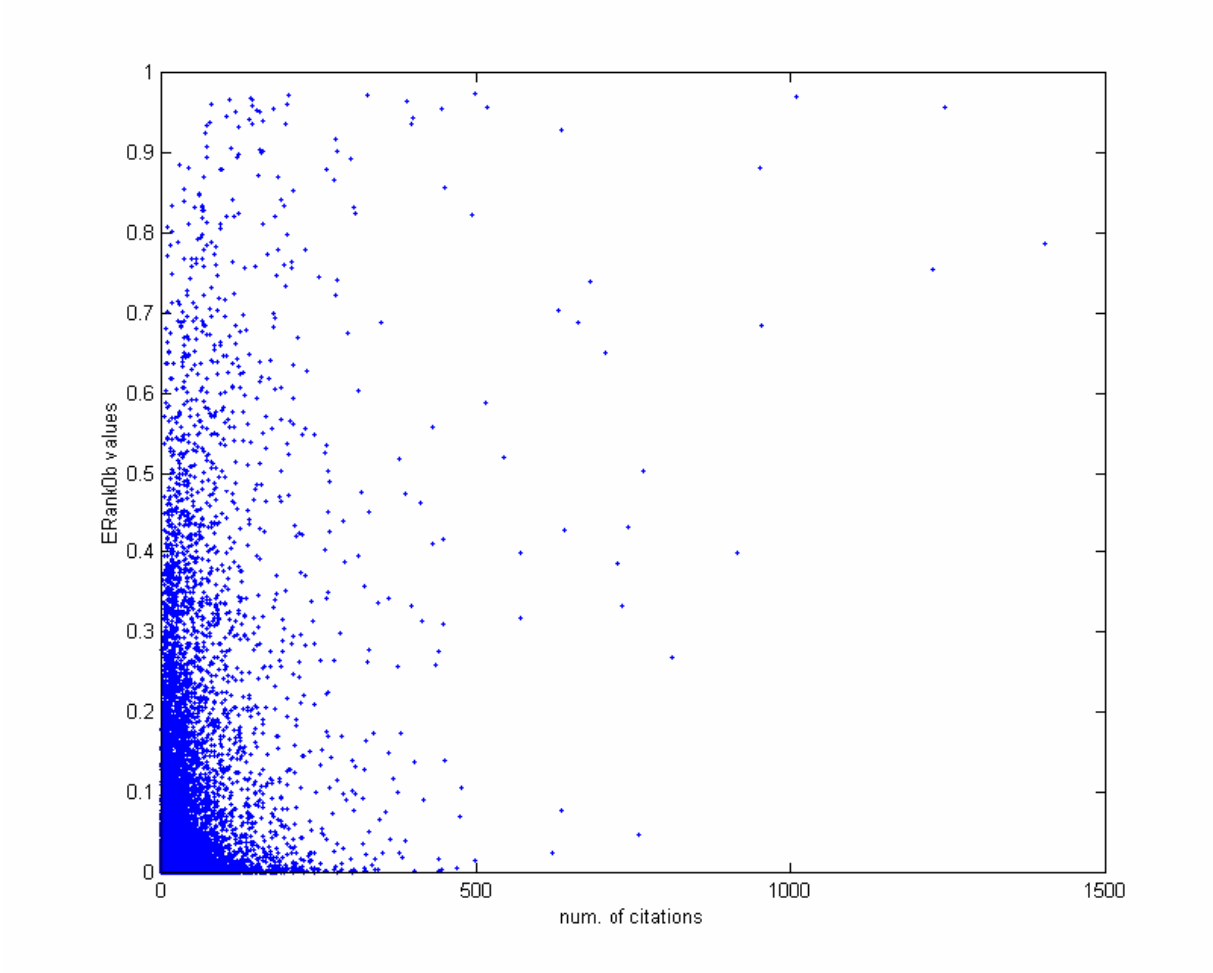

Figure 6.15. Scatter plot of citation count vs. ERank0(b)

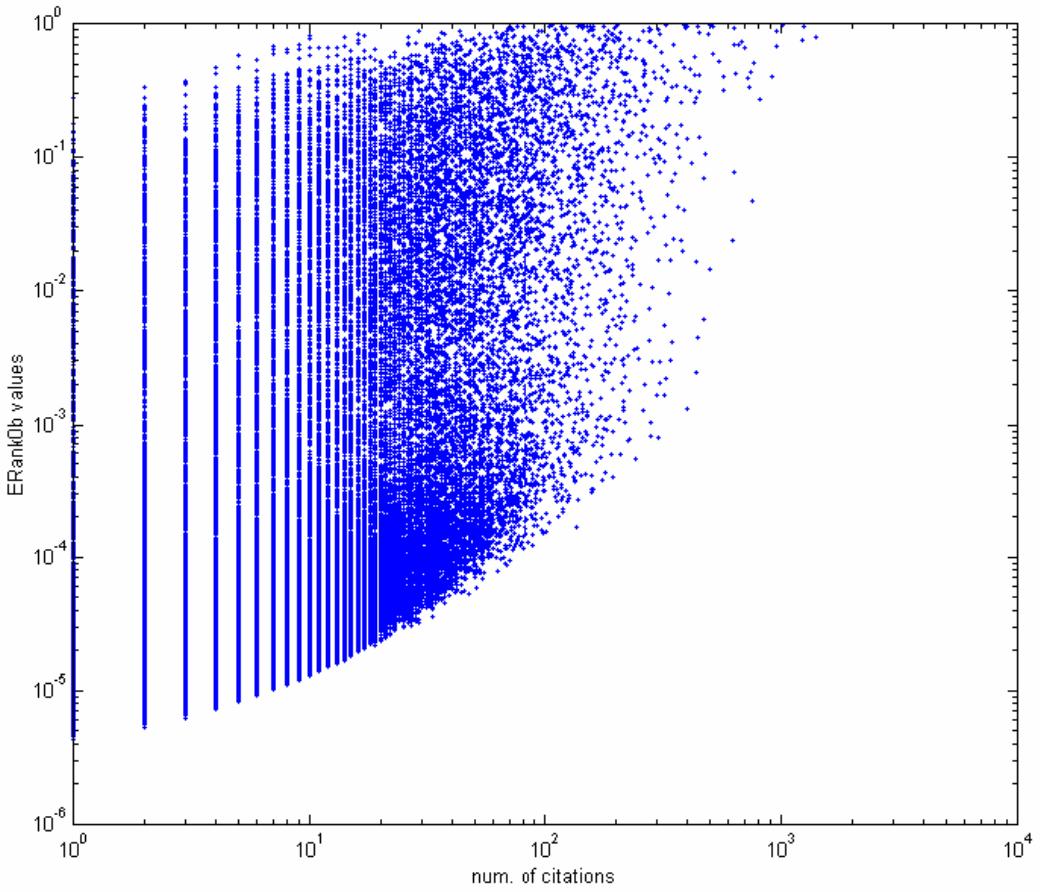

Figure 6.16. Log-log plot of citation count vs. ERank0(b)

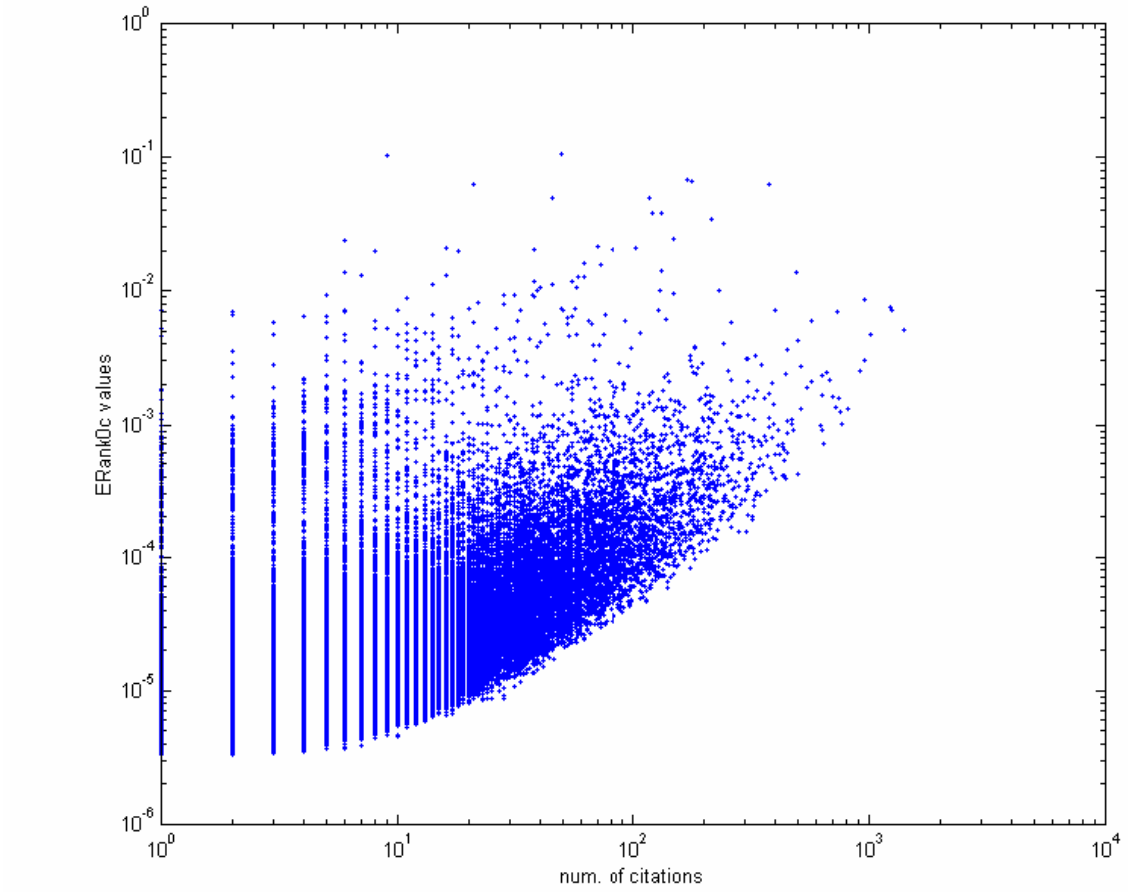

Figure 6.17. Log-log plot of citation count vs. ERank0(c)

6.9.3 ERank0(a) vs. ERank0(b)

The log-log plot between ERank0(a) vs. ERank0(b) (Figure 6.18) helps back our expectation that the relation (a) setting has to (b), is not much stronger than that of CitationCount in the overall sense.

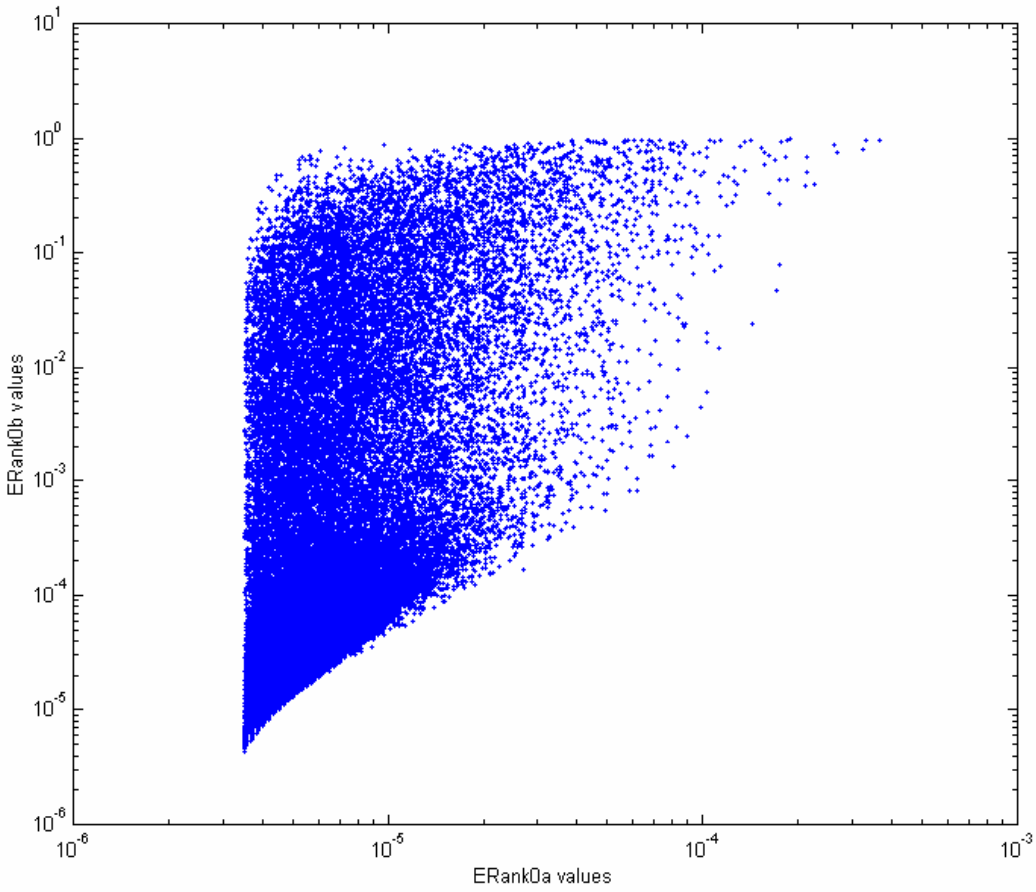

Figure 6.18. Log-log plot of ERank0(a) vs. ERank0(b)

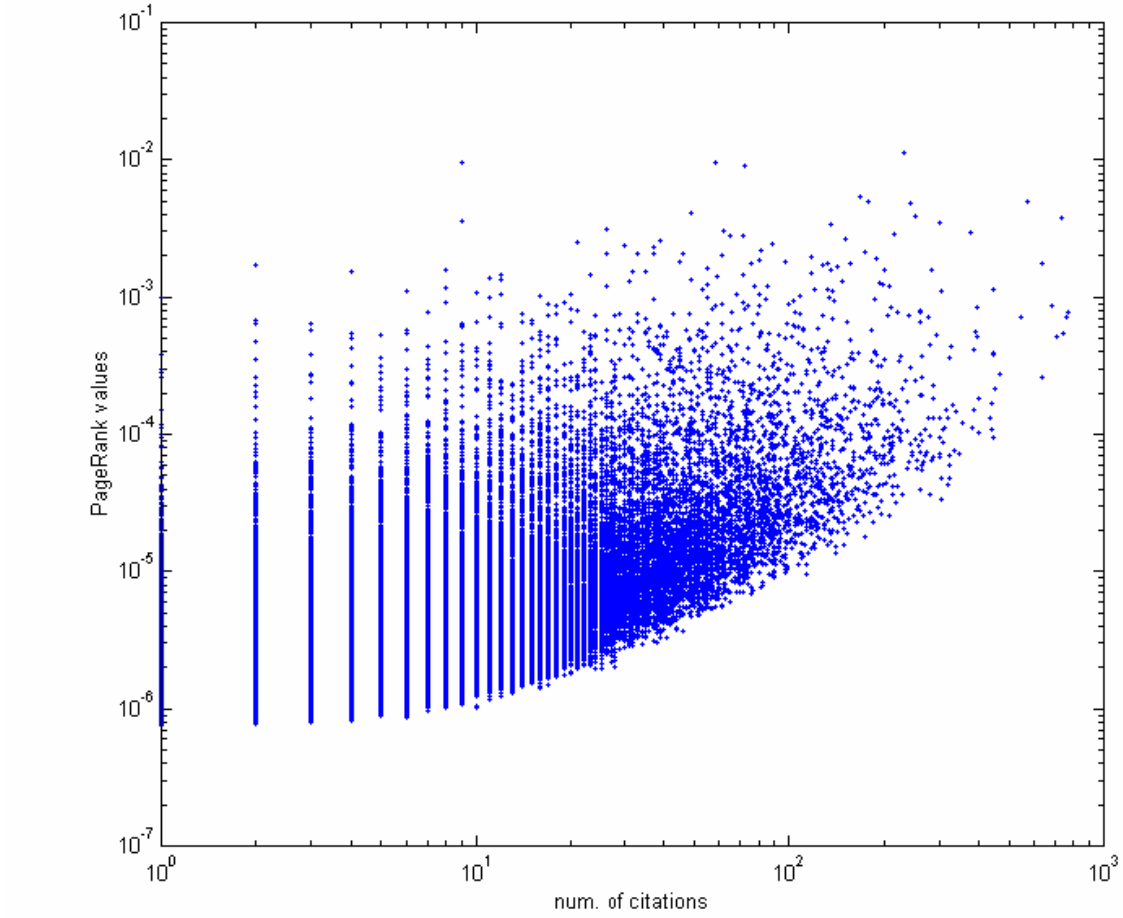

Figure 6.19. Log-log plot of citation count vs. PageRank

6.9.4 CitationCount vs. PageRank

The log-log plot in Figure 6.19 gives an unimpressive relation between CitationCount and PageRank. Visually, it is similar to the (b) and (c) settings. Our assessment on the similarity of PageRank and citation count admits some relation, but it is not a strong one. In this sense our position is closer to (Pandurangan *et al.*, 2002) than (Upstill *et al.*, 2003). Yet, similar to (Pandurangan *et al.*, 2002), we are reluctant to conclude our decision on both PageRank and our other ranking algorithms basing our judgment purely on global properties. As a first step towards understanding community structures we will conduct some query based analysis as demonstrated later in this chapter.

6.9.5 ERank0(c) and ERank0(c2) vs. PageRank

We have created two log-log plots, one for PageRank vs. ERank0(c), and the other one for PageRank vs. ERank0(c2). As suggested by correlation figures on Table 6.5, while the first graph shows some signs of a relation, in the second one a near linear relation is exhibited.

As we have earlier mentioned, ERank0(c2) is a ERank0(c) setting run on the pruned graph for PageRank. The similarity between the two algorithms dramatically diminishes because of the altered structure of the graph after the pruning process. Some 100 000 nodes are discarded in this process, which accounts for about 1/3 of all the nodes. Thus, ERank0(c2) gives a better understanding of the similarity.

As we expect seeing Figure 6.20 using ERank0(c), it is not very obvious what kind of relation to cast on the data. Yet, Figure 6.21 using ERank0(c2) gives a clearer picture.

The line on Figure 6.21 is given as a visual aid, and it marks the linear relationship with a slope 1. Relying solely on visual analysis one can see the linearity of the relationship between the two ranks. Curiously there seems to be a cluster structure with two apparent big components. Being a log-log plot, these point to a dominating linear relation between the ranks, yet the slope of the linearity (as indicated by the vertical position of the clustered points) appears to change within the collection. This may be due to community structures, with different slopes for relating ERank0(c) and PageRank values.

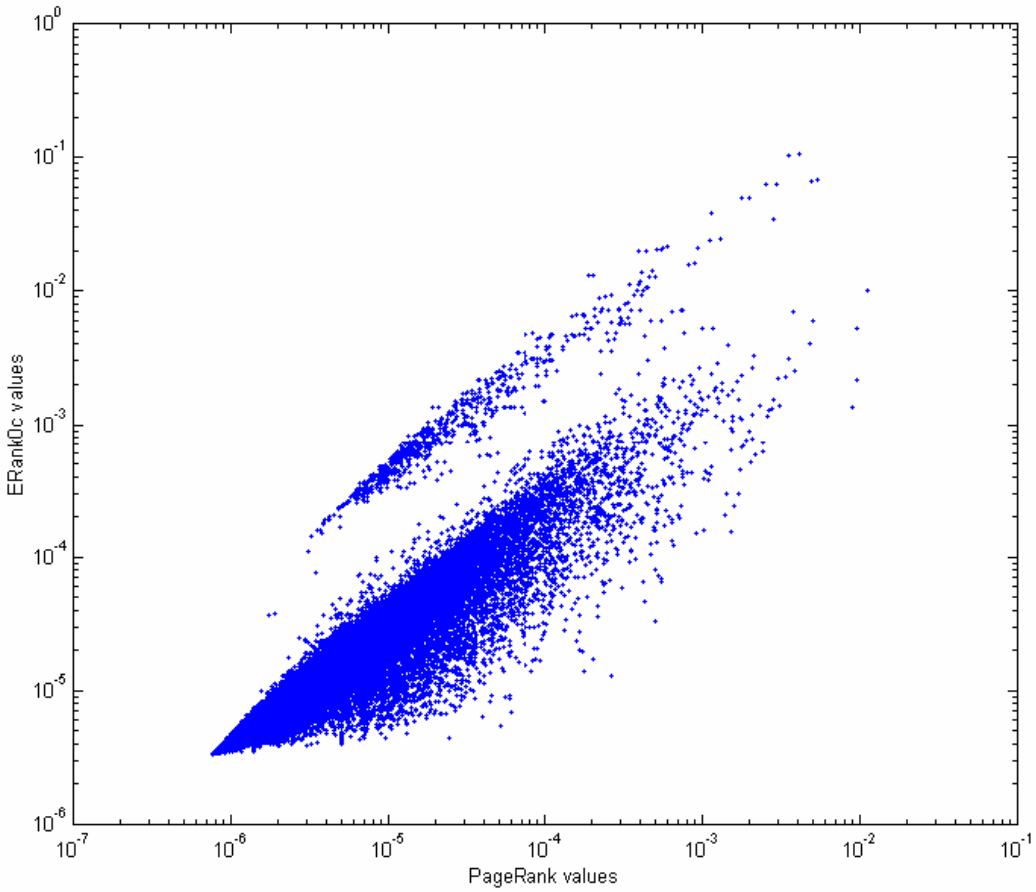

Figure 6.20. Log-log plot of PageRank vs. ERank0(c)

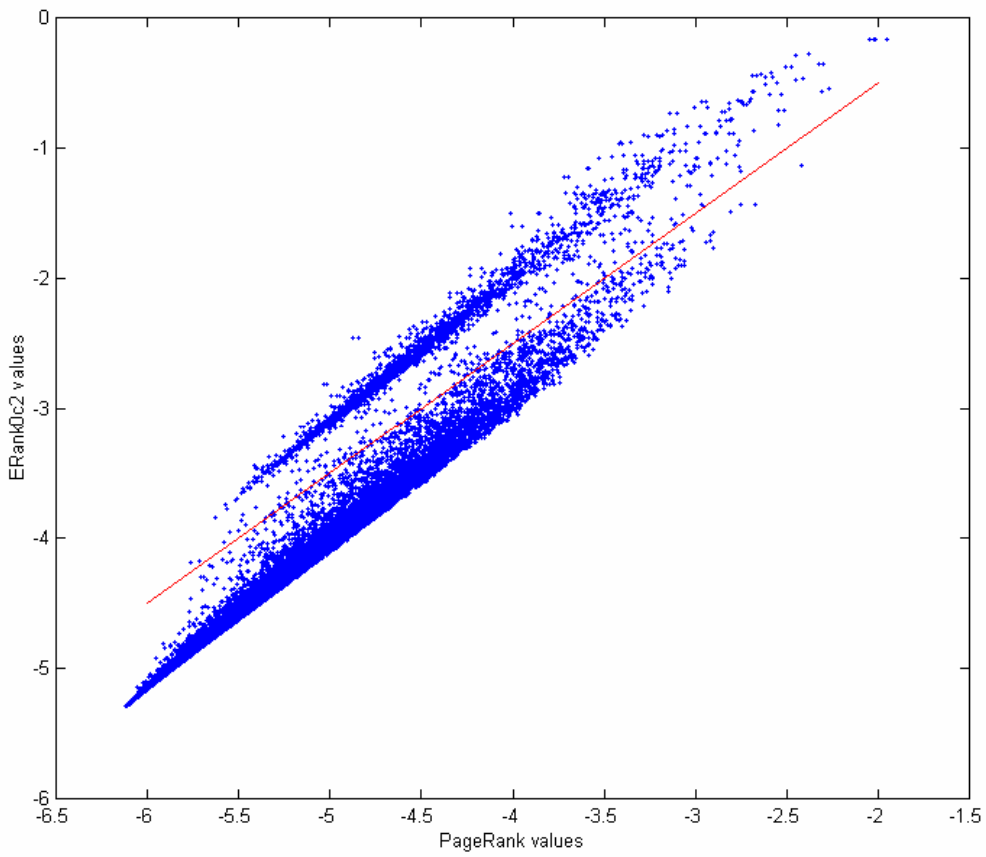

Figure 6.21. Log-log plot of PageRank vs. ERank0(c2)

6.10. Average Position Distance Plots

In this section we introduce a new measure to compare different ranks, which we call average position distance. We compute it by sorting the documents in their rank values and then finding the absolute difference in their positions:

$$APD_{1,2} = \sum_{i=1}^{n} \left| Pos_{1,2}(i) - i \right| \tag{6.3}$$

where *APD* is the average position distance of algorithm 1 with respect to algorithm 2, n gives the number of documents from the top to include with respect to algorithm 2, $Pos_{1,2}$

is a function that maps the position of the i^{th} document with respect to algorithm 2 to its corresponding position (from top) with respect to algorithm 1.

We have found this measure useful for a number of reasons. Firstly, it brings a tangible measure for the similarity of the algorithms having an easily interpretable numerical value. As various ranks assigned to documents range from integers to very small floating point values, the significance of this value can be easier to understand rather than directly comparing the values assigned. It is helpful in the sense that, to some extent it reflects the experience of the searching agent. In this context, it is not the actual rank which matters, rather it is the order of presentation of the documents. Also, it makes it possible to plot the results from all the relevant algorithms in a clean and meaningful plot as we detail below.

We have discovered that, relative distances between algorithms vary from within documents with higher ranks (lower positions) to lower ranking ones as more documents from the top are taken into consideration. To accommodate this phenomenon we have plotted these distances using subsequently greater amount of documents from the top (e.g. top 10, top 100, documents and so on). Each figure contains a log-log plot for top-N documents with regard to the rank being examined.

We have two sets of APD plots. The first set includes ERank0(a), ERank0(b), ERank0(c) and CitationCount algorithms. These algorithms are run on the full network, while the second set of plots is produced on the pruned network as explained on section 5.9. The second set of algorithms are; ERank0(c) and PageRank.

When documents are ordered using the citations counts, ERank0(a) and CitationCount appear to agree largely on importance of documents (i.e. their position from top) when documents have a high amount of citations (Table 6.6). This gradually changes when documents with fewer citations are included as seen in Figure 6.23. This difference may have been amplified by the quasi-random ordering of the documents with fewer citations within the group having the same number of citations. These observations match with the scatter plot of Figure 6.14.

We present in the next section, how these two settings return similar results when filtered using actual queries. The situation is roughly the same also when documents are ordered according to their ERank0(a) values (Figure 6.22). In our interpretation, this suggests that, ERank0(a) has a tendency to favor documents with higher citation values compared to others yet there is a nuance, as it also incorporates global influences.

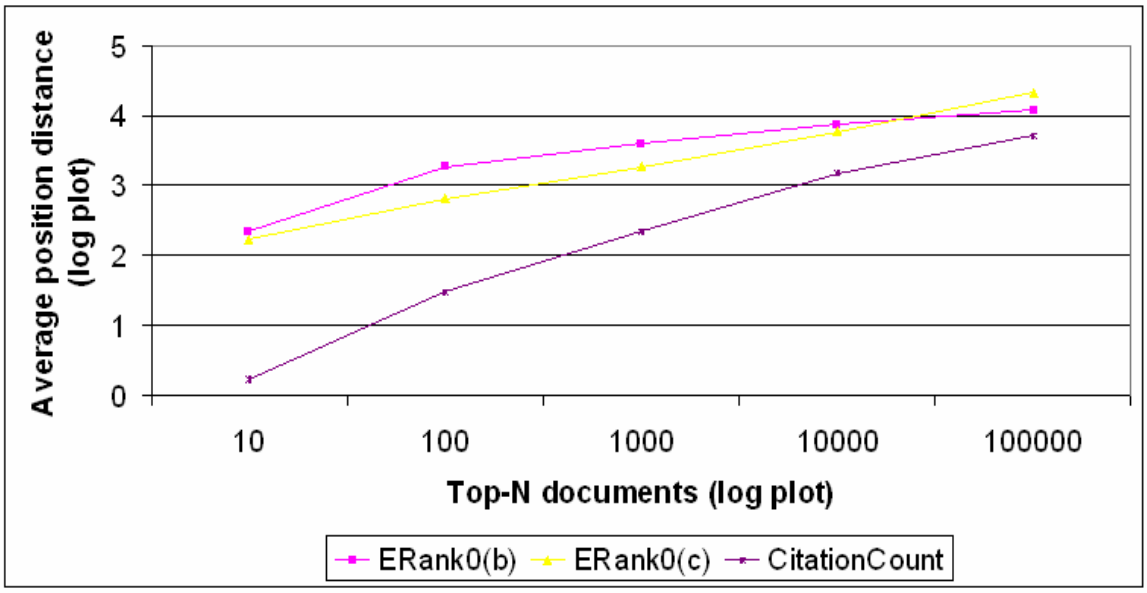

Figure 6.22. Distance plot with respect to ERank0(a)

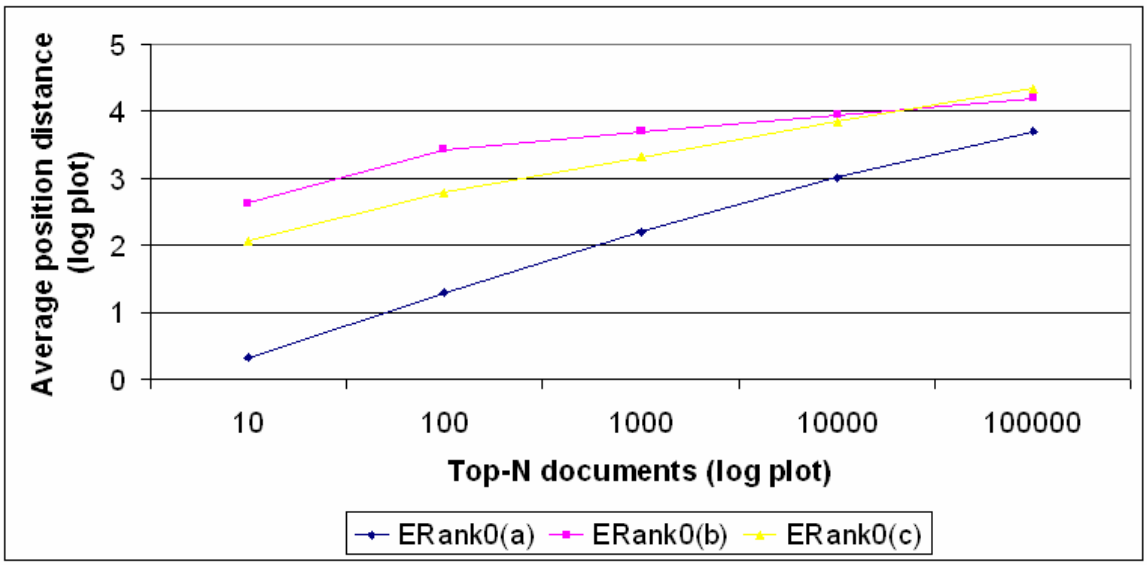

Figure 6.23. Distance plot with respect to CitationCount

Table 6.6. Average position distances for ERank0(a) and CitationCount

Top-N nodes	AvgPos(CitationCount) w.r.t. ERank0(a)	AvgPos(ERank0(a)) w.r.t. CitationCount
10	1.6667	2.1111
100	30.0606	19.0000

1000	225.0050	158.0320
10000	1516.0810	1031.3461
100000	5202.5903	4831.0420

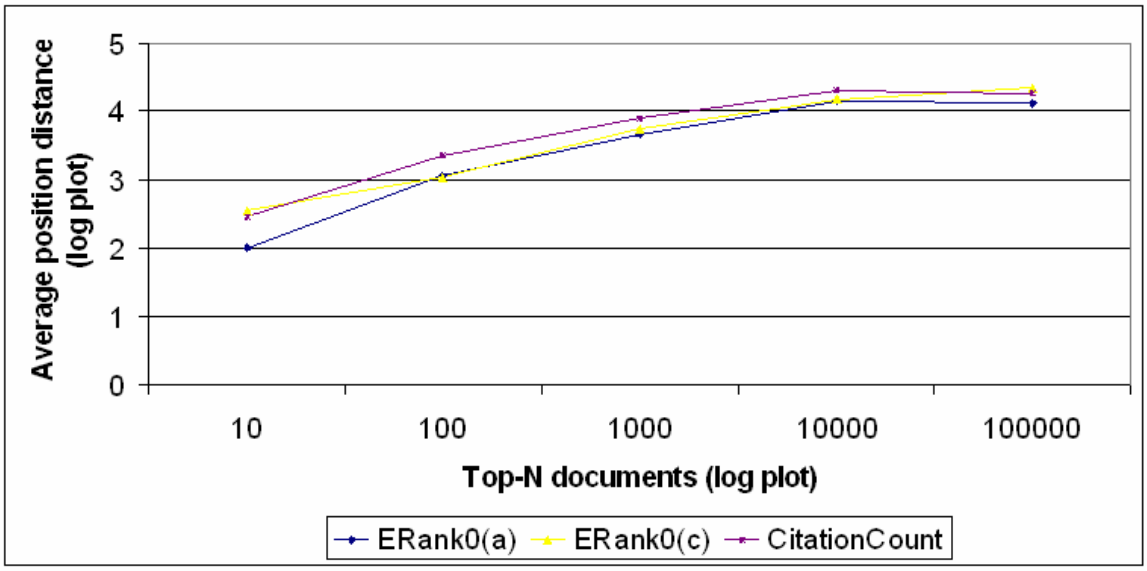

Figure 6.24. Distance plot with respect to ERank0(b)

We see in Figure 6.24 and Figure 6.25 that ERank0(b) and ERank0(c) do not bear any similarity in regard to this measure with the other algorithms.

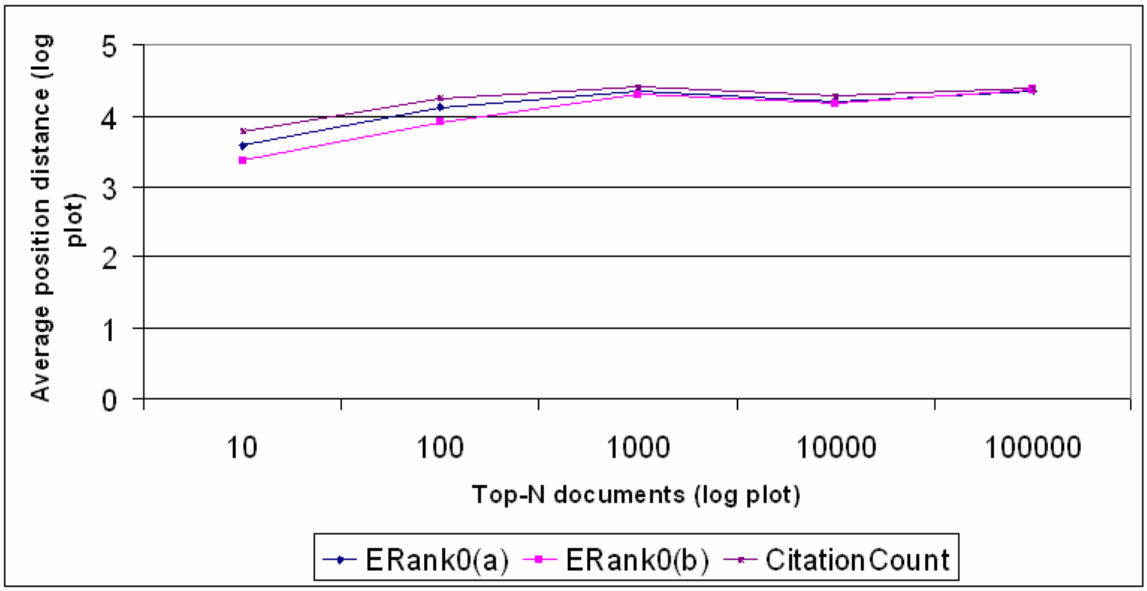

Figure 6.25. Distance plot with respect to ERank0(c)

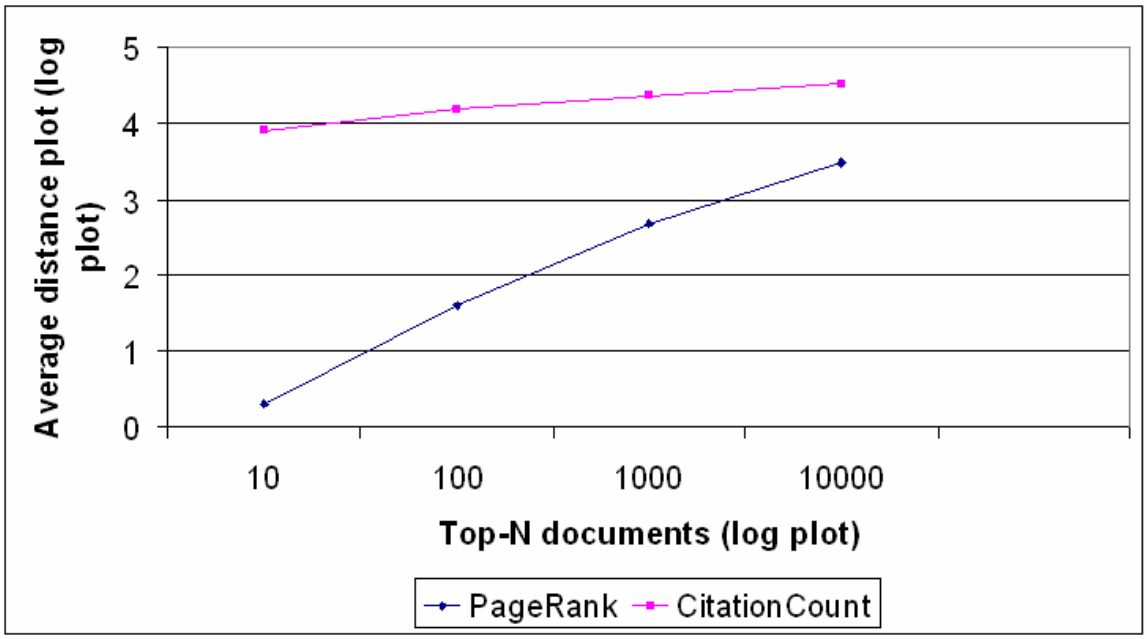

Figure 6.26. Distance plot with respect to ERank0(c2) on the pruned network

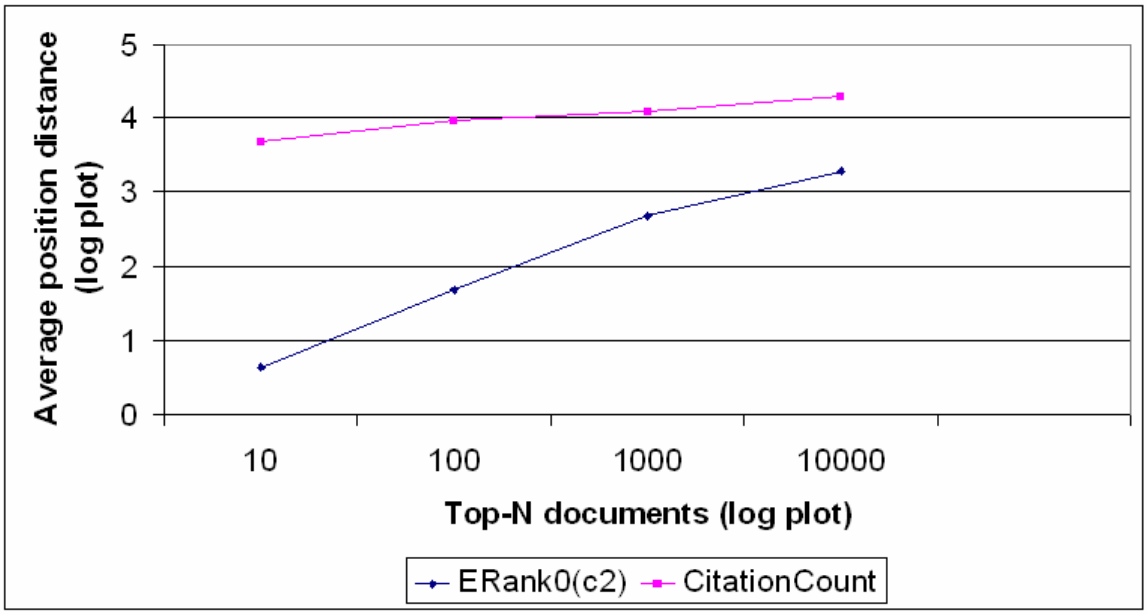

Figure 6.27. Distance plot with respect to PageRank on the pruned network

Table 6.7. Average position distances for ERank0(c2) and PageRank

Top-N nodes	AvgPos(PageRank) w.r.t. ERank0(c2)	AvgPos(ERank0(c2)) w.r.t. PageRank
10	2	4.3333
100	40.7879	47.5051
1000	472.0581	486.1942
10000	3012.8408	1944.6846

Similar to what we see between ERank0(a) and CitationCount, we see a similarity between ERank0(c2) and PageRank (Table 6.7). The log-log plots (Figure 6.20 and Figure 6.21) demonstrate that the ranking positions generated by the two algorithms are decidedly closer to each other, than the CitationCount algorithm also taken into account with them. Although, it is not directly shown here, it would be reasonable to assume that ERank0(a) would also be further away from these two algorithms as well.

6.11. Top Rankers

We believe that the higher ranking documents are important because, they give an idea for the kind of documents the algorithm/settings favor. As a general evaluation of the quality of the whole scientific literature present and rated in this collection is not possible, we opt to displaying the results, and speculate about some the qualities we observe on them.

Table 6.8. Top 20 documents for ERank0(a) ranking

Pos	RecordId	Title	Authors	ERank0a	b pos	c pos	c2 pos	pr pos	cc pos
1	484335	Congestion Avoidance and Control	Jacobson	3.64E-04	4	102			4
2	328445	Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment	Liu, Layland	3.30E-04	12	57			2
3	311874	Graph-Based Algorithms for Boolean Function Manipulation	Bryant	3.24E-04	88	93			1
4	527057	Optimization by Simulated Annealing	Gelatt, Vecchi,Kirkpatrick	2.67E-04	112	49			3
5	28289	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems	Shamir, Adleman,Rivest	2.61E-04	43	46			6
6	49066	Tcl and the Tk Toolkit	Ousterhout	2.28E-04	604	186			7
7	547939	Statecharts: A Visual Formalism for Complex Systems	Harel	2.16E-04	156	157			5
8	4526	Random Early Detection Gateways for Congestion Avoidance	Floyd, Jacobson	2.14E-04	377	525	288	116	9
9	34251	Active Messages: a Mechanism for Integrated Communication and Computation	Eicken, Culler, Goldstein, Schauser	2.13E-04	636	394			13
10	105962	A Scheme for Real-Time Channel Establishment in Wide-Area Networks	Ferrari	1.90E-04	1	117			28
11	309293	High Performance Fortran Language Specification	Message P Forum	1.87E-04	522	309	703	191	11
12	92433	Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism	Clark, Zhang, Shenker	1.86E-04	15	192	371	73	38
13	302704	The Stable Model Semantics For Logic Programming	Gelfond, Lifschitz	1.84E-04	152	188			16
14	19422	Symbolic Model Checking: 10 20 States and Beyond	Dill, Clarke, Burch,Mcmillan,Hwang	1.77E-04	29	204	227	40	18
15	19249	A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing	Floyd, Jacobson, Liu, McCanne, Zhang	1.76E-04	531	831	1622	445	17
16	522243	MPI: A Message-Passing Interface Standard		1.76E-04	1120	399			8
17	25887	Mining Association Rules between Sets of Items in Large Databases	Agrawal, Swami, Imielinski	1.75E-04	3384	297			19
18	55671	Fast Algorithms for Mining Association Rules	Agrawal, Srikant	1.73E-04	4455	416	666	136	10
19	484296	RTP: A Transport Protocol for Real-Time Applications	Schulzrinne, Casner, Frederick, Jacobson	1.70E-04	188	305	857	200	14
20	411542	RSVP: A New Resource ReSerVation Protocol	Zhang, Deering, Estrin, Shenker, Zappala	1.63E-04	136	601			20

Table 6.9. Top 20 documents for ERank0(b) ranking

Pos	RecordId	Title	Authors	ERank0b	a pos	c pos	c2 pos	pr pos	cc pos
1	105962	A Scheme for Real-Time Channel Establishment in Wide-Area Networks	Verma, Ferrari	0.973129	10	117			28
2	15039	Non-Deterministic Exponential Time Has Two-Prover Interactive Protocols	Babai,Fortnow,Lund	0.971816	128	294	628	316	255
3	422908	Symbolic Model Checking for Real-time Systems	Henzinger,Nicollin,Sifakis, Yovine	0.971183	39	479	410	243	79
4	484335	Congestion Avoidance and Control	Jacobson	0.969496	1	102			4
5	17094	A Really Temporal Logic	Alur, Henzinger	0.968038	174	226	90	45	530
6	522428	From ATP to Timed Graphs and Hybrid Systems	Sifakis, Yovine	0.966757	192	528			918
7	2915	Algebraic Methods for Interactive Proof Systems	Lund, Fortnow	0.96668	233	327	672	366	522
8	463585	Proof Verification and the Hardness of Approximation Problems	Arora, Lund, Motwani, Sudan, Szegedy	0.963882	46	232	598	183	53
9	124	Hybrid Automata: An Algorithmic Approach to the Specification and Verification of Hybrid Systems	Alur, Courcoubetis, Henzinger, Pei-Hsin Ho	0.960752	131	966	1359	690	266
10	58451	On the Power of Multi-Prover Interactive Protocols	Rompel, Sipser	0.959721	469	379	636	357	1584
11	54530	An Implementation of the Contract Net Protocol Based on Marginal Cost Calculations	Sandholm	0.958056	452	2272			513
12	328445	Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment	Liu, Layland	0.957441	2	57			2
13	142710	Learning to Predict by the Methods of Temporal Differences	Sutton	0.956901	26	173			25
14	251152	A Spatial Logic based on Regions and Connection	Randell, Cui, Cohn	0.955515	472	1822	2800	960	357
15	92433	Supporting Real-Time Applications in an Integrated Services Packet Network: Architecture and Mechanism	Clark, Zhang, Shenker	0.954405	12	192	371	73	38
16	222622	Designing Programs That Check Their Work	Blum, Kannan	0.951978	250	333	657	348	471
17	131548	Automatic Symbolic Verification of Embedded Systems	Pei-hsin Ho, Alur, Henzinger	0.95173	214	1807	1937	1190	453
18	70445	Real-time Logics: Complexity and Expressiveness	Alur, Henzinger	0.950289	355	292	106	54	768
19	155016	Sender-Based Message Logging	Johnson, Zwaenepoel	0.94573	580	805			978
20	524648	Implementing Remote Procedure Calls	Birrell, Nelson	0.942658	27	56			50

Table 6.10. Top 20 documents for ERank0(c) ranking

Pos	RecordId	Title	Authors	ERank0c	a pos	b pos	c2 pos	pr pos	cc pos
1	516071	Probabilistic Methods in Combinatorics	Spencer	0.105287	2011	367	5	9	3620
2	239544	Discrepancy in Arithmetic Progressions	Spencer	0.103844	24467	2050	6	12	31427
3	549100	Structure and Complexity of Relational Queries	Harel, Chandra	0.067168	184	95	21	5	388
4	548351	Computable Queries for Relational Databases	Harel, Chandra	0.066632	247	139	23	7	354
5	20336	Generalized Additive Models	Hastie, Tibshirani	0.06336	81	6076	25	17	59
6	284366	Hazard Regression	Kooperberg, Stone, Truong	0.062528	5227	8736	26	23	12455
7	93436	Privacy Enhancement for Internet Electronic Mail: Part II: Certificate-Based Key Management	Kent	0.049396	824	1271	30	34	787
8	219414	Privacy Enhancement for Internet Electronic Mail: Part III: Algorithms, Modes, and Identifiers	Balenson	0.049032	1176	765	33	37	4074
9	130506	Set-Oriented Production Rules in Relational Database Systems	Widom, Finkelstein	0.038172	485	2033	62	75	750
10	93250	Deriving Production Rules for Constraint Maintenance	Widom, Ceri	0.03817	522	2029	63	74	618
11	35316	Relational Queries Computable in Polynomial Time	Immerman	0.03441	151	174	51	18	209
12	340970	An Empirical Study of Learning Speed in Back-Propagation Networks	Fahlman	0.024528	745	3216	61	62	498
13	222340	Accelerated Learning in Back-Propagation Nets	Fahlman	0.024194	22072	5950	65	76	43716
14	74056	On Serializability Of Multidatabase Transactions Through Forced Local Conflicts	Georgakopoulos, Rusinkiewicz, Sheth	0.021304	1718	5112	118	164	1989
15	228254	Model Selection and Accounting for Model Uncertainty in Linear Regression Models	Raftery, Madigan, Hoeting	0.021197	790	2807	134	93	995
16	185576	2PC Agent Method: Achieving Serializability In Presence Of Failures In A Heterogeneous Multidatabase	Wolski, Veijalainen	0.021153	6988	5807	120	181	16992
17	225770	Unveiling Turbo Codes: Some Results on Parallel Concatenated Coding Schemes	Benedetto, Montorsi	0.020676	2208	15511	158	187	1523
18	229736	Design of Parallel Concatenated Convolutional Codes	Benedetto, Montorsi	0.020567	5760	17006	159	206	5339
19	161628	Relativized Counting Classes: Relations among Thresholds, Parity, and Mods	Beigel	0.020141	12877	1214	73	242	14609
20	514357	On The Relativized Power of Additional Accepting Paths	Beigel	0.019948	26696	1101	74	286	36875

Table 6.11. Top 20 documents for CitationCount ranking

Pos	RecordId	Title	Authors	Citation Count	a pos	b pos	c pos	c2 pos	pr pos
1	311874	Graph-Based Algorithms for Boolean Function Manipulation	Bryant	1404	3	88	93		
2	328445	Scheduling Algorithms for Multiprogramming in a Hard-Real-Time Environment	Liu, Layland	1244	2	12	57		
3	527057	Optimization by Simulated Annealing	Kirkpatrick, Gelatt, Vecchi	1225	4	112	49		
4	484335	Congestion Avoidance and Control	Jacobson	1010	1	4	102		
5	547939	Statecharts: A Visual Formalism for Complex Systems	Harel	955	7	156	157		
6	28289	A Method for Obtaining Digital Signatures and Public-Key Cryptosystems	Shamir, Adleman,Rivest	952	5	43	46		
7	49066	Tcl and the Tk Toolkit	Ousterhout	917	6	604	186		
8	522243	MPI: A Message-Passing Interface Standard	Message P Forum	813	16	1120	399		
9	4526	Random Early Detection Gateways for Congestion Avoidance	Floyd, Jacobson	766	8	377	525	288	116
10	55671	Fast Algorithms for Mining Association Rules	Agrawal, Srikant	759	18	4455	416	666	136
11	309293	High Performance Fortran Language Specification		742	11	522	309	703	191
12	552631	Fast Anisotropic Gauss Filtering	Geusebroek, Smeulders, Weijer	733	21	849	62	113	11
13	34251	Active Messages: a Mechanism for Integrated Communication and Computation	von Eicken, Culler, Goldstein, Schauser	727	9	636	394		
14	484296	RTP: A Transport Protocol for Real-Time Applications	Schulzrinne, Casner, Frederick, Jacobson	705	19	188	305	857	200
15	86872	Rewrite Systems	Jouannaud,Dershowitz	680	23	119	222	470	102
16	302704	The Stable Model Semantics For Logic Programming	Gelfond, Lifschitz	661	13	152	188		
17	19249	A Reliable Multicast Framework for Light-weight Sessions and Application Level Framing	Floyd, Jacobson, Liu, McCanne, Zhang	639	15	531	831	1622	445
18	19422	Symbolic Model Checking: 10 20 States and Beyond	Burch, Clarke, McMillan, Dill, Hwang	635	14	29	204	227	40
19	25887	Mining Association Rules between Sets of Items in Large Databases	Agrawal, Imielinski,Swami	635	17	3384	297		
20	411542	RSVP: A New Resource ReSerVation Protocol	Zhang, Deering, Estrin, Shenker, Zappala	630	20	136	601		

Table 6.12. Top 20 documents for ERank0(c2) ranking on the pruned graph

Pos	RecordId	Title	Authors	ERank0c	pr pos	cc pos
1	148879	Yacc: Yet Another Compiler-Compiler	Johnson	0.010187074	1	179
2	400432	Pilot: An Operating System for a Personal Computer	Redell et.al.	0.002168122	3	2667
3	42081	Lint, a C Program Checker	Johnson	0.005209681	2	33206
4	242933	Experience with Processes and Monitors in Mesa	Lampson,Redell	0.001335569	4	1913
5	516071	Probabilistic Methods in Combinatorics	Spencer	0.105287281	9	3620
6	239544	Discrepancy in Arithmetic Progressions	Spencer	0.103843915	12	31427
7	22491	Finding Structure in Time	Elman	0.005910474	6	22
8	149498	A Learning Algorithm for Continually Running Fully Recurrent Neural Networks	Williams, Zipser	0.004009068	8	161
9	221670	Improving Register Allocation for Subscripted Variables	Callahan, Carr, Kennedy	0.002299918	14	588
10	225305	Blocking Linear Algebra Codes For Memory Hierarchies	Carr, Kennedy	0.001387816	15	9000
11	390693	The Mutual Exclusion Problem - Part I: A Theory of Interprocess Communication	Lamport	0.001143524	22	5164
12	386741	The Mutual Exclusion Problem Part II: Statement and Solutions	Lamport	0.001372849	26	5534
13	74862	Memory-Efficient Algorithms for the Verification of Temporal Properties	Courcoubetis, Vardi, Wolper, Yannakakis	0.000590638	27	1530
14	254456	State-Space Caching Revisited	Godefroid, Holzmann, *et al.*	0.000509306	32	6588
15	60865	Cryptographic Limitations on Learning Boolean Formulae and Finite Automata	Kearns, Valiant	0.001289654	21	480
16	128963	Training A 3-Node Neural Network Is NP-Complete	Blum, Rivest	0.000616038	24	1311
17	94985	Efficient Induction Of Logic Programs	Feng, Muggleton	0.002540587	10	154
18	222642	Inductive Logic Programming	Muggleton	0.003111988	13	101
19	16747	Dynamic Parameter Encoding for Genetic Algorithms	Schraudolph,Belew	0.001326157	20	2277
20	531412	Delta Coding: An Iterative Search Strategy for Genetic Algorithms	Whitley, Mathias, Fitzhorn	0.000715378	25	7582

Table 6.13. Top 20 documents for PageRank ranking on the pruned graph

Pos	RecordId	Title	Authors	PageRank	c2 pos	cc pos
1	148879	Yacc: Yet Another Compiler-Compiler	Johnson	0.011208	1	101
2	42081	Lint, a C Program Checker	Johnson	0.009613	3	22659
3	400432	Pilot: An Operating System for a Personal Computer	Redell et.al.	0.009531	2	1828
4	242933	Experience with Processes and Monitors in Mesa	Lampson,Redell	0.008962	4	1271
5	549100	Structure and Complexity of Relational Queries	Harel, Chandra	0.005353	21	235
6	22491	Finding Structure in Time	Elman	0.005025	7	9
7	548351	Computable Queries for Relational Databases	Harel, Chandra	0.004951	23	212
8	149498	A Learning Algorithm for Continually Running Fully Recurrent Neural Networks	Williams, Zipser	0.004772	8	91
9	516071	Probabilistic Methods in Combinatorics	Spencer	0.004153	5	2457
10	94985	Efficient Induction Of Logic Programs	Feng Muggleton	0.003894	17	86
11	552631	Fast Anisotropic Gauss Filtering	Geusebroek, Smeulders, Weijer	0.003803	113	4
12	239544	Discrepancy in Arithmetic Progressions	Spencer	0.003534	6	21233
13	222642	Inductive Logic Programming	Muggleton	0.003512	18	49
14	221670	Improving Register Allocation for Subscripted Variables	Callahan, Carr, Kennedy	0.003358	9	360
15	225305	Blocking Linear Algebra Codes For Memory Hierarchies	Carr, Kennedy	0.003098	10	6501
16	324526	Reaching Approximate Agreement in the Presence of Faults	Dolev, Lynch, Pinter, Stark	0.003035	41	1672
17	20336	Generalized Additive Models	Hastie, Tibshirani	0.002973	25	29
18	35316	Relational Queries Computable in Polynomial Time	Immerman	0.002856	51	120
19	566858	A New Fault-Tolerant Algorithm for Clock Synchronization	Lundelius, Lynch	0.002832	40	1302
20	16747	Dynamic Parameter Encoding for Genetic Algorithms	Schraudolph,Belew	0.002796	19	1526

Firstly, we note that ERank0(a) yields results with the highest citation counts, so these are already the documents we have come to expect as highly regarded. Similarly, we see that the top-20 results returned by ERank0(c2) and PageRank are highly similar.

In ERank0(b) we see that, a good ratio of the top ranks are dominated by groups of authors (Herzinger, Alur, Lund, ...). Likewise, many of the papers appear to be about what might loosely be termed “formal verification” and “theorem proving”. Comparing with the results of (a), we interpret this as suggesting that with ERank0(b) a whole community of papers seems have been lifted by a global domination of ranks. We contrast this with (a) and citation count, in which we have greater variety of papers whose ranks rest more on local influences, although with (a) – as indicated by its shift from citation count – the communities still probably do exist.

In ERank0(c), we see a striking similarity to PageRank, and the results in top 20 are very similar. In both ERank0(c) and PageRank we see multiple papers from the same author appear together in the top 20.

One fact is that, most of the papers ranked highly are dating from early 1990’s or earlier. This shows that, in order to be ranked very highly, documents need time to have developed a network of citations surrounding them.

Overall, we note that even for the top 20 documents the ranking algorithms are not in good agreement. It is not possible to see from these results, if this is caused by differences between the global rankings of communities the papers belong to, or do the ranking algorithms also favor papers in a similar topic differently as well?

To investigate this question further on, in the following section we present numerous query results using different ranking schemes.

6.12. Sample Query Results

In this section, we present the results from two queries in detail out of the ones we have run on our collection. The key words are searched in the title, author summary (names, affiliation etc.), and description (approx. first 1000 characters) fields for each document. We show the different rankings produced by different schemes.

The unavailability of some of the most influential papers in the related topics coupled with the absence of also their citation data must unfortunately have compromised the quality of the ranks computed, along with leaving gaps within the result sets returned.

We have selected queries such that the results are a subset selected from a larger set of query hits available, and with higher number of citation counts, so that the effects of the different ranking schemes can be observed. We retained, and display the number of citations as a conventional measure of the information value of a paper.

Running other queries on topics with fewer citation counts – less "dense" areas – in our data set, we have observed that the results returned appear simply to follow the citation count order. This may happen both due to under representation of the topic in the dataset (e.g. "small world"), or simply the lack of publications on the field.

ERank0(a) setting appears largely to follow citation count order, occasionally altering the order with a few steps amongst returned documents.

ERank0(b) can make dramatic changes to the results (as compared to their citation count ordering), and it may be favoring papers from authors who have a collection of highly cited papers.

The similarity of the results returned by ERank0(c) and PageRank are also evident here. On our numerous query runs the returned results were highly similar in the documents returned and their ordering. This actually is not surprising, recalling the average position distance figures on Table 6.7.

Our general impression has been that, query based analysis is much likely to give a valuable insight on the structure of the network. Yet the samples presented here only give an example for the prospective insight to be gained.

Query: "*dempster shafer*"
Number of hits: 74 and 47 (pruned network)

Table 6.14. Top 10 query results sorted using ERank0(a) ranks

Query rank	Title	Author	Citation Count	(b) pos	(c) pos	(cc) pos
1	A Logic for Reasoning about Probabilities	Halpern, Fagin	86	1	2	1
2	Numerical Uncertainty Management in User and Student Modeling: An Overview of Systems and Issues	Jameson	31	5	5	2
3	Rough Mereology: A New Paradigm For Approximate Reasoning	Skowron	25	9	8	3
4	Cluster-based Specification Techniques in Dempster-Shafer Theory	Schubert	15	11	13	4
5	A New Approach to Updating Beliefs	Halpern, Fagin	11	4	1	5
6	Quantitative Modeling of User Preferences for Plan Recognition	Bauer	10	6	15	6
7	Logic-based Plan Recognition for Intelligent Help Systems	Paul, Bauer	8	3	10	9
8	Some qualitative approaches to applying the Dempster-Shafer theory	Parsons	10	17	33	6
9	Representing and Retrieving Structured Documents using the Dempster-Shafer Theory of Evidence: Modelling and Evaluation	Ruthven, Lalmas	9	18	24	8
10	Possibilistic Semantics and Measurement Methods in Complex Systems	Joslyn	8	13	16	9

Table 6.15. Average position distances for Top-10 results w.r.t. ERank0(a) results

	ERank0(a)	ERank0(b)	ERank0(c)	CitationCount
ERank0(a)	0	4.2	8	0.6
ERank0(b)	4.2	0	4.8	4.8
ERank0(c)	8	4.8	0	8.2
CitationCount	0.6	4.8	8.2	0

Table 6.16. Top 10 query results sorted using ERank0(c2) ranks

Query rank	Title	Author	Citation Count	(pr) pos	(cc) pos
1	A Logic for Reasoning about Probabilities	Halpern, Fagin	86	1	1
2	A New Approach to Updating Beliefs	Halpern, Fagin	11	2	3
3	Finding A Posterior Domain Probability Distribution By Specifying Nonspecific Evidence	Schubert	7	3	9
4	A Hybrid Framework for Representing Uncertain Knowledge	Saffiotti	4	4	15
5	A Hybrid Belief System For Doubtful Agents	Saffiotti	1	6	36
6	Possibilistic Semantics and Measurement Methods in Complex Systems	Joslyn	8	8	5
7	A Defect in Dempster-Shafer Theory	Wang	6	7	10
8	Numerical Uncertainty Management in User and Student Modeling: An Overview of Systems and Issues	Jameson	31	5	2
9	Logic-based Plan Recognition for Intelligent Help Systems	Paul, Bauer	8	10	5
10	A Dempster-Shafer Approach to Modeling Agent Preferences for Plan Recognition	Bauer	6	11	10

Table 6.17. Average position distances for Top-10 results w.r.t. ERank0(c2) ordering

	ERank0(c2)	PageRank	CitationCount
ERank0(c2)	0	0.8	6.3
PageRank	0.8	0	6.3
CitationCount	6.3	6.3	0

Query: "*information retrieval*"

Number of hits: 1309, 895 (pruned network)

Table 6.18. Top 10 query results sorted using ERank0(a) ranks

Query rank	Title	Author	Citation Count	(a) pos	(b) pos	(cc) pos
1	Information Retrieval	Rijsbergen	473	8	2	1
2	Querying the World Wide Web	Mendelzon, Mihaila	210	22	11	2

3	Mobile Agents: Are They a Good Idea?	Chess, Harrison, Kershenbaum	175	46	17	3
4	Searching Distributed Collections With Inference Networks	Callan, Lu, Croft	152	56	23	4
5	Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays	Ahlberg, Shneiderman	151	60	8	5
6	Automatic Resource list Compilation by Analyzing Hyperlink Structure and Associated Text	Chakrabarti, Dom, Gibson, Keinberg, Raghavan, Rajagopalan	110	51	27	8
7	Using Linear Algebra for Intelligent Information Retrieval	Berry, Dumais, O'Brien	132	57	22	6
8	NewsWeeder: Learning to Filter Netnews	Lang	95	41	26	11
9	SIFT - A Tool for Wide-Area Information Dissemination	Yan	102	31	13	9
10	Affective Computing	Picard	126	91	7	7

Table 6.19. Average position distances for Top-10 results w.r.t. ERank0(a) ordering

	ERank0(a)	**ERank0(b)**	**ERank0(c)**	**CitationCount**
ERank0(a)	0	40.8	10.7	0.9
ERank0(b)	40.8	0	30.7	40.7
ERank0(c)	10.7	30.7	0	10
CitationCount	0.9	40.7	10	0

Table 6.20. Top 10 query results sorted using ERank0(c2) ranks

Query rank	**Title**	**Author**	**Citation Count**	**(pr) pos**	**(cc) pos**
1	Subtopic Structuring for Full-Length Document Access	Hearst	61	1	16
2	The Effectiveness of Navigable Information Disclosure Systems	Bosman, Bouwman, Bruza	14	5	121
3	The Modelling and Retrieval of Documents using Index Expressions	Bruza, Weide	12	6	146
4	How May I Help You?	Gorin, Riccardi, Wright	39	8	30
5	Distributed Indexing: A Scalable Mechanism for Distributed Information Retrieval	Danzig, Ahn, Noll, Obraczka	24	2	73
6	Visual Information Seeking: Tight Coupling of Dynamic Query Filters	Ahlberg, Shneiderman	151	3	2

	with Starfield Displays				
7	Applications of Approximate Word Matching in Information Retrieval	Powell, French	6	51	279
8	Algorithms for Scoring Coreference Chains	Bagga, Baldwin	6	53	279
9	Affective Computing	Picard	126	4	3
10	Automatic Routing and Ad-hoc Retrieval Using SMART : TREC 2	Buckley, Salton, Allan	21	7	80

Table 6.21. Average position distances for Top-10 results w.r.t. ERank0(c2) ordering

	ERank0(c2)	**PageRank**	**CitationCount**
ERank0(c2)	0	11.3	89.3
PageRank	11.3	0	102.8
CitationCount	89.3	102.8	0

7. DISCUSSION AND CONCLUSION

7.1. A Review of Work Done

In this work we have introduced and analyzed a framework we named PAS-ETRI for analyzing complex networks using Probabilistic Argumentation Systems (PAS). PAS-ETRI offers a generic way to map a graph structure to a corresponding PAS instance.

An ETRI can be used to analyze a variety of complex networks like the Web, citations networks and biological networks amongst others. Various aspects of networks can be analyzed depending on the PAS-ETRI model designed such as relevance (of documents) or ranking (of web pages) or community structures (of authors) amongst others.

We have focused on two models; Document Relevance Model (DRM) and Document Information Value Model (DIM). DRM deals with the relevance problem, whilst DIM is used for ranking documents which has been our main focus. As a fundamental concept for ranking we have introduced Minimal Evidence (ME) which mimics maximum-likelihood (ML) hypothesis for maximum a posteriori (MAP) learning. Using DIM and ME we define ArgRank, which involves computations for an NP-hard problem. We intend to use PAS-ETRI based ranking for very large networks, so this algorithm is not suitable. Yet ArgRank is based using clear semantics on well established evidential reasoning techniques.

This brought us to the second main theme of this work, which is applying PAS-ETRI in an efficient manner using approximation algorithms. We have introduced a novel family of algorithms, which we have named ETRI Support Propagation (ESP). ESP is based on the common conjunction model of a network. It is applicable for networks in which neighboring nodes in a network share a fairly constant amount of common conjunction in their supporting arguments which is represented by a damping function. This becomes a valid assumption for ETRI models in which the link assumptions have a

low value. The zeroth order ESP algorithm ESP-0 is susceptible to feedback from neighboring nodes, thus ESP-0 based algorithms produce vulnarable results on networks with many cross-links. ESP-1 on the other hand deals with this problem, by using a message passing algorithm instead. ESP algorithms, and ESP-1 in particular are similar in spirit to Belief Propagation (BP) algorithm of Pearl. (Pearl, 1988)

We have made particular emphasis on the analysis of ESP-0 algorithm, and present a theoretical analysis, using which we reveal that under certain conditions ESP-0 produces non-decreasing results (per iteration) which are bound from above by the true dsp values. ESP-0 is inferior to ESP-1, thus these results have an indirect implication for ESP-1 as well. We define ERank-0 as a straight-forward application of ESP-0 to ranking, using a constant damping function.

We have presented various experimental results. We used a scientific citation network using data from the CiteSeer network (CIT) which contained about 300 000 nodes and 1 250 000 directed links. We have run three different algorithms for comparison; ERank-0, citation count (in-degree), and PageRank. ERank-0 can produce dramatically different results depending on the way link assumption probabilities are assigned. We have run three different settings; (a) and (b) use constant values, (c) uses variable values inversely proportional to the out-degree of a node.

We have initially studied the rank distributions produced by our various algorithms. For citation count we have confirmed the generally accepted power-law distribution with an exponent 3.0 (Redner, 1998) (Redner, 2004) (Newman, 20003). We have found a power-law distribution for PageRank values as in (Pandurangan *et al.*, 2002). For (a) and (c) settings the power-law holds, whereas for (b) it did not.

Our study reveals that a main characteristic in defining ranking algorithms is how they balance local versus global influences. Citation count represents the local extremum in this sense, whereas ERank-0 algorithms can be parameterized to lie on a wide range. In this sense we identify PageRank as a globally dominated ranking algorithm, which can highly disagree with citation count.

We introduce and use a measure we call average position distance (APD), and use it for generating comparison plots. These are helpful because APD provides a way to compare all the ranking algorithms on the same basis; the ranks they create, which yields easy to interpret and tangible results. Our APD plots have checked with our scatter plots for disclosing similarities between the algorithms.

We have presented top ranking documents as a way to demonstrate results favored by different algorithms. Similarly we have presented example query results along with corresponding APD plots. These were helpful in showing concrete examples for the agreements and disagreements of ranking algorithms for actual usage from the point of a cognitive agent.

7.2. Discussion and Directions for Theoretical Aspects

We hope our work to stimulate interest in evidential reasoning techniques for analyzing complex networks. The ETRI framework provides a generic way to this, and what we have presented in this work is a limited picture of the possible uses.

A variety of uses for PAS-ETRI can be formulated. An immediate such use as future work for us is its use as a tool for detecting community structures in a social network (or topics in a scientific citation/author network). It can also be used as an analysis tool where more fine grained results are needed, as PAS offers a systematic way to deal with “detail”. It can be anticipated this list can be extended to many systems where a network based modeling has made sense.

We believe, for Information Retrieval PAS based approaches may have an important potential. (Picard and Savoy, 2003) We do our part in this work by introducing the ESP family of algorithms which make it possible to apply PAS-ETRI based algorithms to very large networks with acceptable accuracy.

For an effective IR scheme, the rank merging problem is one of the foremost issues to be addressed. In this sense, a more general application combining DRM and DIM could create a far more effective search engine, this remains as an important future work.

As we have earlier mentioned PageRank builds on a very similar network structure like DIM. Thus extensions to PageRank create a natural direction future work for further research on this topic (Richardson and Domingos, 2001) (Ingongngam and Rungsawang, 2003) (Haveliwala, 2002) (Kao *et al.*, 2002).

In this work we have mainly focused on a general ranking scheme. However, it is well possible to alter it to include personalization and specialization. In this sense, these can be perceived as simply incorporation of extra "evidence" to the network structure in addition to the minimal evidence (ME).

The common conjunction model presents a clear way of modeling and simplifying relations between nodes in a network structure. While it has experimentally proven to be useful in our applications, its relationship with complex network features such as clustering and degree distributions is a topic to be addressed. In relation, it would be beneficial to study the damping function value distributions as a complex network characteristic. In this context its relation to generative models of complex networks could prove to be useful, and may be employed in generating more accurate models.

The ESP algorithms we have presented have had only two orders 0 and 1. Although we have covered an extent of the properties of ESP-0 there is yet left to be done. Whilst we have shown a worst-case convergence property for ESP-0 using the lowest-only estimates for damping values, the actual usage is not like this. Our initial experimentation using incrementally increasing damping values and examining the top ranking documents (which we did not show here) has shown that the deviation from results by using different damping values is not very significant. Yet we believe this remains a fact to be established.

Also, we have discovered for some settings a "*first-diverge then converge*" behavior in experimentations (also not shown in this work). This is possibly due to the capping of

possible dsp values by 1.0 (as they are probabilities) from above, so an indefinite divergence behavior is not possible anyway. This remains an attitude of our algorithms to be examined.

Our study of ESP family has fallen short of presenting a detailed study of the properties of ESP-1, focusing instead on ESP-0. We consider an implementation and experimentation of ESP-1 along with a theoretical treatment as a direct and necessary follow-up for our work.

Once the theoretical implications of ESP-1 is revealed, it may prove useful to introduce higher order ESP-n algorithms, partially employing n^{th} order calculations for revealing the micro-structure and using the support propagation paradigm for the global picture. This may prove to be a very useful tool against link spamming and manipulation because it would have virtually no algorithmic weaknesses for such manipulation up to n^{th} order.

We have used a constant damping value for our experimentations. However, this is not the only option available. A damping function employing some heuristics may yet emerge to yield better approximation results, especially on networks where the common conjunction model with a constant value is not a good representative. For example a network with highly variable clustering properties with frequently high link assumption probabilities would be difficult to analyze using a constant damping function.

7.3. Discussion and Directions for Experimental Results and Methodology

We have used the CiteSeer citation network as our main data. Occasionally we have also used scale-free random networks for experimentation. Our choice of a citation network over a web sample had pros and cons. Firstly, we believe that a citation network constitutes a more significant and more important network structure, because references on a paper represent a much intenser study behind and thus are by any means a stronger evidence of a real relationship. We are building all our effort on this aspect of the network; that it ultimately encodes evidence, or relationship information. So the more the quality of

evidence present, the better our algorithms should work, and thus a citation network is a better demonstration bed for the usefulness of our algorithms than a Web sample. Also, related to the former, is that the Web has grown very complicated in its link structure. Dynamically produced documents (as opposed to static HTML pages of 90's) have complex page and link structures, a significant amount of commercial ads, and manipulation by link spammers which create an essentially polluted link structure, and thus present an additional challenge towards adapting the application of any link algorithm to distill any results. (see for example (Kao *et al.*, 2002))

Yet ultimately the goal of our algorithms are to be applicable to all sorts of networks, and the Web is one of the biggest – if not the biggest – of the complex network. Ranking algorithms certainly have great and important prospects for use in search engines, and so we believe that an application and evaluation of our algorithms to a Web structure is an important follow-up.

We have experimented with various link assumption probability assignment schemes. A study focused on this very aspect should prove useful. An establishment of the conventional assignments for different complex network types (e.g. citation networks, the Web), and possibly new schemes would prove very beneficial.

An interesting product of our study was to reveal that some ERank-0's produce power-law distributions while others may not. The character of generating and deviating form a power-law distribution and its relation in affecting actual results would be a very interesting relationship to disclose.

Our results revealed that ERank-0 results can be highly correlated to citation count, as well as PageRank. While we have presented a good deal of results on this, we believe there is still more to be done on this, possibly evaluating different types of networks and link assumption probability schemes.

The similarity of ERank0(c) and PageRank may deserve extra attention. As their mathematical formulations suggest, their similarity is more than experimental, but that

actually PageRank under certain assumptions can be interpreted as an approximate form of ERank0(c) or vice-versa. This actually, may prove useful in shedding light on the success of PageRank from an evidential reasoning perspective, and remains as an interesting future work for us.

Our discussions have essentially been on the characteristics of ranking algorithms, on their balancing local and global influences. There are other LAR algorithms, and an examination of them in this perspective should bring an interesting insight for their workings.

While we have presented some example query results, and have run and examined numerous such, it has been out of scope of this work due to time and space limitations to include a comprehensive query-based analysis. It remains a desirable future work to do a systematic treatment of the subject, for example running many queries from a controlled keyword repository of a publisher and examine the results. Such an analysis could be a key to disclosing the inherent community structures in a citation network, and most important of all would show us the effect of different ranking schemes on intra- and inter-community ranks. We perceive the work of (Upstill *et al.*, 2003) as one initial effort towards such a query based analysis. This analysis would bring us an important insight which is not possible to attain examining a general and overall picture.

APPENDIX A. PROOFS

A.1. Proof of Theorem 2.1

1. Clearly,

$$a \hat{\vee} b = 1-(1-a)(1-b) = b \hat{\vee} a$$

2. Using the definition;

$$\begin{aligned} a \hat{\vee} (b \hat{\vee} c) &= 1-(1-a)(1-b \hat{\vee} c) \\ &= 1-(1-a)\left[1-\left(1-(1-b)(1-c)\right)\right] = 1-(1-a)(1-b)(1-c) \\ &= 1-\left[1-\left(1-(1-a)(1-b)\right)\right](1-c) = 1-(1-a \hat{\vee} b)(1-c) \\ &= (a \hat{\vee} b) \hat{\vee} c \end{aligned}$$

□

A.2. Proof of Theorem 2.2

Using Ineq.(2.32) we get:

$$a \hat{\vee} b \quad \geq \quad a$$

$$\left(a \hat{\vee} b\right) - a \quad \geq \quad 0$$

Let us define:

$$\begin{aligned} \Delta &= (a \hat{\vee} b) - a \\ &= \left(1-(1-a)(1-b)\right) - a \\ &= \left(1-\left(1-a-b+ab\right)\right) - a \end{aligned}$$

$$= (a + b - ab) - a$$
$$= b - ab$$
$$= b(1 - a)$$

As we know $0 \le a \le 1$ and $0 \le b \le 1$, we see:

$$(1 - a) \ge 0$$

So, it follows for Δ:

$$\Delta = b(1 - a) \ge 0$$

This proves that inequality(2.32) holds, concluding the proof.

□

A.3. Proof of Theorem 3.1

We will outline a constructive proof here using the path finding paradigm. We can simply write the support for a node as follows:

$$SP(v_i, \xi) = a_i \vee \bigvee_{k=2}^{k\max} T_k(v_i, \xi)$$

where $T_k(v_i, \xi) \subseteq SP(v_i, \xi)$ is a set representing the disjunction of all the supporting terms of a vertex v_i of the k^{th} order (i.e. with k literals), and $kmax$ is the order of the longest supporting argument of the give node.

Re-writing it more explicitly and recalling the path-finding process for finding the supporting arguments give us:

$$SP(v_i,\xi) = a_i \vee$$

$$\left[\left(\bigvee_{v\in P_i(0)} a_v \wedge l_{vi}\right) \vee \left(\bigvee_{v\in P_i(1)}\bigvee_{w\in P_v(0)} \left(a_w \wedge l_{wv} \wedge l_{vi}\right)\right) \vee \left(\bigvee_{v\in P_i(1)}\bigvee_{w\in P_v(1)}\bigvee_{x\in P_w(0)} \left(a_x \wedge l_{xw} \wedge l_{wv} \wedge l_{vi}\right)\right) \vee \ldots\right]$$

where $P_i(0) \subseteq P_i$ denotes parents of *i* which have no parents, and $P_i(1) \subseteq P_i$ nodes which have parents. Re-arranging it gives us:

$$SP(v_i,\xi) = a_i \vee$$

$$\left[\left(\bigvee_{v\in P_i(0)} l_{vi} \wedge a_v\right) \vee \left(\bigvee_{v\in P_i(1)} l_{vi} \wedge \bigvee_{w\in P_v(0)} \left(l_{wv} \wedge a_w\right)\right) \vee \left(\bigvee_{v\in P_i(1)} l_{vi} \wedge \bigvee_{w\in P_v(1)} l_{wv} \wedge \bigvee_{x\in P_w(0)} \left(l_{xw} \wedge a_x\right)\right) \vee \ldots\right]$$

$$SP(v_i,\xi) = a_i \vee \left[\bigvee_{v\in P_i} l_{vi} \wedge \left\{a_v \vee \left[\left(\bigvee_{w\in P_v(0)} \left(l_{wv} \wedge a_w\right)\right) \vee \left(\bigvee_{w\in P_v(1)} l_{wv} \wedge \bigvee_{x\in P_w(0)} \left(l_{xw} \wedge a_x\right)\right) \vee \ldots\right]\right\}\right]$$

Note here how the expression inside the set parenthesis is actually the support for vertex *v* $SP(v_v,\xi)$.

$$SP(v_i,\xi) = a_i \vee \left[\bigvee_{v\in P_i} l_{vi} \wedge SP(v_v,\xi)\right]$$

This gives us the equation of Theorem 3.1 as desired.

□

A.4. Lemma A.1

Given real numbers $0 \le a_1, a_2, ..., a_n < 1,\ 0 \le b_1, b_2, ..., b_n < 1,\ 0 \le c_1, c_2, ..., c_n \le 1$ *and:*

$$\hat{\bigvee_{i=1..n}} a_i \ge \hat{\bigvee_{i=1..n}} b_i$$

Then it follows that:

$$\hat{\bigvee}_{i=1..n} c_i a_i \geq \hat{\bigvee}_{i=1..n} c_i b_i$$

A.5. Proof of Lemma A.1

Using the given equation with Eq.(2.30):

$$\hat{\bigvee}_{i=1..n} a_i \geq \hat{\bigvee}_{i=1..n} b_i$$

$$1-\prod_{i=1}^{n}(1-a_i) \geq 1-\prod_{i=1}^{n}(1-b_i)$$

$$\prod_{i=1}^{n}(1-b_i) \geq \prod_{i=1}^{n}(1-a_i)$$

We can get log of both sides recalling $a_i \neq 1,\ b_i \neq 1$:

$$\log\left(\prod_{i=1}^{n}(1-b_i)\right) \geq \log\left(\prod_{i=1}^{n}(1-a_i)\right)$$

$$\sum_{i=1}^{n}(1-b_i) \geq \sum_{i=1}^{n}(1-a_i)$$

$$\sum_{i=1}^{n}(a_i-b_i) \geq 0$$

$$\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b_i \geq 0$$

$$\sum_{i=1}^{n} a_i \geq \sum_{i=1}^{n} b_i$$

Thus given $c_i \neq 0$ (to avoid getting log of 0), to make the proof we need to show that:

$$\sum_{i=1}^{n}(c_i a_i) \geq \sum_{i=1}^{n}(c_i b_i)$$

given

$$\sum_{i=1}^{n} a_i \geq \sum_{i=1}^{n} b_i$$

We can simply divide each side by c_i's on both sides giving the required inequality. For the case when $c_i = 0$, we note that 0 has no effect on the results as in:

$$a \hat{\vee} 0 = a$$

So, we can simply remove the members of the series where $c_i = 0$, and proceed with the proof.

A.6. Proof of Theorem 5.1

We will use induction to prove our theorem.

We will deal here with the case where $d\hat{s}p_i(s) \neq 1$ for any *s*, but this case can be shown to hold similarly.

BASIS STEP:

We will show that the following inequality holds:

$$\forall v_i \in V : d\hat{s}p_i(1) \geq d\hat{s}p_i(0)$$

Note that $d\hat{s}p_i(0)$ represents the initial value assignment by the algorithm. So $d\hat{s}p_i(0)$ is a set with *n* elements such that:

$$d\hat{s}p_i(0) = \{0, 0, .., 0\}$$

Using equation (5.7) we get:

$$\begin{aligned} d\hat{s}p_i(1) &= p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(0) \right] \\ &= p(a_i) \hat{\vee} \left[d_c(v_i) \cdot 0 \right] \\ &= p(a_i) \end{aligned}$$

We know by definition that $p(a_i) \geq 0$.

INDUCTIVE STEP:

We will assume that the following inequality holds for any *s*:

$$\forall v_i \in V : d\hat{s}p_i(s) \geq d\hat{s}p_i(s-1)$$

Using Eq.(5.7) we get:

$$d\hat{s}p_i(s) = p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s-1) \right]$$

We also know that:

$$d\hat{s}p_i(s+1) = p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s) \right]$$

Let us define Δ such that:

$$\Delta = d\hat{s}p_i(s+1) - d\hat{s}p_i(s)$$

We can write:

$$\begin{aligned}
\Delta &= d\hat{s}p_i(s+1) - d\hat{s}p_i(s) \\
&= p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s) \right] - p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s-1) \right] \\
&= d_c(v_i) \left[\hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s) - \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s-1) \right] \\
&= d_c(v_i) \left[\hat{\bigvee}_{j \in P_i} r_j(s) - \hat{\bigvee}_{j \in P_i} r_j(s-1) \right]
\end{aligned}$$

where we define $r_j(s) = p(l_{ji}) d\hat{s}p_j(s)$. So, we get:

$$\Delta \propto \left[\hat{\bigvee}_{j \in P_i} r_j(s) - \hat{\bigvee}_{j \in P_i} r_j(s-1) \right]$$

But using Lemma A.1 on Eq.(5.7) we can show that:

$$d\hat{s}p_i(s) \geq d\hat{s}p_i(s-1) \text{ implies } \left[\hat{\bigvee}_{j \in P_i} r_j(s) \geq \hat{\bigvee}_{j \in P_i} r_j(s-1) \right]$$

This, in turn implies that:

$$\Delta \geq 0$$

which shows that $\forall v_i \in V : d\hat{s}p_i(s+1) \geq d\hat{s}p_i(s)$.

Thus, by induction we have proved our initial assumption, this concludes the proof.

□

A.7. Proof of Theorem 5.2

We will deal here with the case where $dsp_i \neq 1$ and $d\hat{s}p_i(s) \neq 1$ for any *s*, but this case can be shown to hold similarly.

Firstly, we are given the following equality by the theorem.

$$\forall v_i \in V : dsp_i \geq 1 - (1 - p(a_i)) d_c(v_i) \prod_{j \in P_i} (1 - p(l_{ji}) dsp_j)$$

We can re-write the inequality using the noisy-or operator:

$$\forall v_i \in V : dsp_i \geq p(a_i) \hat{\vee} d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) dsp_j$$

We will use induction to prove our theorem.

BASIS STEP:

We know that the following inequality holds.

$$\hat{dsp}_i(0) \leq dsp_i$$

as we know that $\hat{dsp}_i(0) = \{0, 0, .., 0\}$ by definition of the ESP-0 algorithm.

We will show that this implies

$$\hat{dsp}_i(1) \leq dsp_i$$

Using Eq.(5.7) we know that:

$$\begin{aligned} \hat{dsp}_i(1) &= p(a_i) \hat{\vee} d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) \hat{dsp}_j(0) \\ &= p(a_i) \leq dsp_i \end{aligned}$$

which provides our basis step.

INDUCTIVE STEP:

Let us assume now the following equation holds.

$$\hat{dsp}_i(s) \leq dsp_i \tag{A.1}$$

We will show that this implies:

$$\hat{dsp}_i(s+1) \leq dsp_i$$

Let us first define:

$$dsp_i' = p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) dsp_j \right] \tag{A.2}$$

so by the theorem we are ensured that:

$$dsp_i' \leq dsp_i$$

We also know that:

$$d\hat{s}p_i(s+1) = p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s) \right] \tag{A.3}$$

Using Lemma A.1 and Ineq. (A.1) we can show:

$$\hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s) \leq \hat{\bigvee}_{j \in P_i} p(l_{ji}) dsp_j$$

which after some manipulation becomes:

$$p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} p(l_{ji}) d\hat{s}p_j(s) \right] \leq p(a_i) \hat{\vee} \left[d_c(v_i) \hat{\bigvee}_{j \in P_i} \left(p(l_{ji}) dsp_j \right) \right]$$

Using Eq.(A.2) and Eq.(A.3) on this we get:

$$d\hat{s}p_i(s+1) \leq dsp_i'$$

But using Eq.(A.1) we get:

$$d\hat{s}p_i(s+1) \le dsp'_i \le dsp_i$$

$$d\hat{s}p_i(s+1) \le dsp_i$$

Thus, by induction we conclude the proof.

□

A.8. Proof of Theorem 5.3

We will use proof by contradiction. Let us assume now that the algorithm runs indefinitely. Then for any $s > 1$ we have:

$$difference_i(s) > \delta_i$$

so,

$$difference_i(s) > e_0$$

Using Eq. (5.4) we can get:

$$d\hat{s}p_i(s) > d\hat{s}p_i(s-1) + e_0$$

We see that:

$$d\hat{s}p_i(1) > d\hat{s}p_i(0) + e_0 = e_0$$

$$d\hat{s}p_i(2) > d\hat{s}p_i(1) + e_0 > 2 \cdot e_0$$

$$\vdots$$

$$d\hat{s}p_i(s) > s \cdot e_0$$

and,

$$s < \frac{d\hat{s}p_i(s)}{e_0} \leq \frac{1}{e_0}$$

recalling that $0 \leq d\hat{s}p_i(s) \leq 1$. This inequality shows that s has to remain a finite value, which contradicts with our initial assumption. Thus by contradiction we proved that the algorithm terminates after a finite number of iterations.

□

APPENDIX B. A BDD BASED PAS-ETRI IMPLEMENTATION

In this appendix we will present a brief overview of the PAS-ETRI implementation we have created. Within the context of this work, we have used it to approximately calculate damping constants, and compare results from ERank-0 outputs.

Our implementation has been capable of handling mostly around 100-200 supporting arguments when the argument order is higher than 2. This number is influenced by how the nodes and links in the network are arranged. This has made it possible for us to examine support with arguments up to the 5^{th} order on our CiteSeer citation network.

For this work, we leave aside a proper introduction of BDDs and related technical aspects of our application, as this in itself is a wide topic which deserves a dedicated treatment and would diverge our focus from approximating PAS-ETRI results. We leave this as a future work. Targeting those who are familiar with the topic we will essentially relate some important choices we have made, their justifications, and the results. The interested reader can consult (Antoine *et al.*, 2003) for a targetted treatment of the use of BDDs for a similar purpose (i.e. calculating the probability for a sum-of-products formula).

We have chosen to use a static variable ordering for our BDD implementations as opposed to dynamic ordering. This is mainly because, in our preliminary research it was possible to find various highly efficient and stable open source implementations for this type of BDD, whilst the use of dynamic variable ordering was not available. A useful prospect of this use is the ability to operate on the support of different nodes, thus allowing various different schemes (comparing, joining, ...) for analyzing complex network properties.

The immediately following pertinent issue is the ordering of the variables within the BDD. A significant amount of research appears to have been made on this topic, and there are various heuristics developed in the literature. We have used the variable ordering created by a depth-first path finding algorithm with a limited depth (corresponding to the

desired maximum argument order) incrementally assigning the variable order for each node met. This sufficed for the creation of a reasonably powerful system serving our purpose, yet much remains to be done on the topic for establishing related facts.

Our algorithm recursively explores the graph backwards on the links starting from the target node using the limited depth first search scheme. A supporting argument which is a product of the literals (on the path) is revealed this way in each step, and is represented by a corresponding BDD. This product is added to the support of the node, thus effectively creating the sum-of-products propositional sentence which is also a BDD. Once this sentence is obtained, the probability of the sentence is computed by traversing the BDD downwards (towards the terminal nodes 1 and 0) on the nodes of the BDD as in (Antoine *et al.*, 2003). This way of calculating the probability is faster than adding the probabilities of individual disjoint terms.

In our implementation we have used an iterative approach in which we have incremented the order limit starting from 2 up to 5, thus obtaining the best possible result. The analysis aborts when a limit is exceeded for the number of the nodes of the BDD (e.g. 2 000 000 nodes), which is determined by the RAM available.

We have used the open source BDD implementation in Java called JavaBDD (JBD), using which through the native interface we have employed the CUDD package (CUD) written in C. The running time for disclosing the support of a node could go up to the order of 1000 seconds.

REFERENCES

(Antoine *et al.*, 2003) Antoine, R., Chatelet, E., Dutuit, Y., and Berenguer, C. (2003). A practical comparison of methods to assess sum-of-products. *Reliability Engineering and System Safety*, 79:33–42.

(Borodin *et al.*, 2005) Borodin, A., Roberts, G. O., Rosenthal, J. S., and Tsaparas, P. (2005). Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297.

(Brin and Page, 1998)Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *WWW7: Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

(Broder *et al.*, 2000a) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000a). Graph structure in the Web. In *Proceedings of the 9th international World Wide Web conference on Computer networks : the international journal of computer and telecommunications netowrking*, pages 309–320, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co.

(Broder *et al.*, 2000b) Broder, A. Z., Krauthgamer, R., and Mitzenmacher, M. (2000b). Improved classification via connectivity information. In *SODA '00: Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 576–585, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

(Bryant, 1986)Bryant, R. E. (1986). Graph-based algorithms for Boolean function manipulation. *IEEE Trans. Comput.*, 35(8):677–691.

(Chen *et al.*, 2004) Chen, Y.-Y., Gan, Q., and Suel, T. (2004). Local methods for estimating PageRank values. In *CIKM '04: Proceedings of the thirteenth ACM conference*

on Information and knowledge management, pages 381–389, New York, NY, USA. ACM Press.

(CIT) CiteSeer database http://citeseer.ist.psu.edu

(Cozman, 2000) Cozman, F. G. (2000). Credal networks. *Artif. Intell.*, 120(2):199–233.

(CUD) CUDD BDD package http://vlsi.colorado.edu/ fabio/cudd

(Dempster, 1968) Dempster, A. P. (1968). A generalization of Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247.

(Ding *et al.*, 2002) Ding, C., He, X., Husbands, P., Zha, H., and Simon, H. D. (2002). Pagerank, HITS and a unified framework for link analysis. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–354, New York, NY, USA. ACM Press.

(Haenni, 2003) Haenni, R. (2003). Detecting conflict-free assumption-based knowledge bases. In *Intelligent Systems for Information Processing: From Representation to Applications*, pages 203–209, North-Holland.

(Haenni *et al.*, 2000) Haenni, R., Kohlas, J., and Lehmann, N. (2000). Probabilistic Argumentation Systems. In Kohlas, J. and Moral, S., editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Volume 5: Algorithms for Uncertainty and Defeasible Reasoning*, pages 221–287. Kluwer, Dordrecht.

(Handcock *et al.*, 2003) Handcock, M. S., Jones, J. H., and Morris, M. (2003). On "sexual contacts and epidemic thresholds," models and inference for sexual partnership distributions. Technical report, arxiv.org.

(Haveliwala, 1999) Haveliwala, T. H. (1999). Efficient computation of PageRank. Technical Report 1999-31, Stanford Digital Library Technologies Project.

(Haveliwala, 2002) Haveliwala, T. H. (2002). Topic-sensitive PageRank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA. ACM Press.

(Heckerman and Breese, 1996) Heckerman, D. and Breese, J. (1996). Causal independence for probability assessment and inference using Bayesian networks. *IEEE Systems,Man, and Cybernetics*, 26:826–831.

(Heidtmann, 1989) Heidtmann, K. (1989). Smaller sums of disjoint products by subproduct inversion (KDH). *IEEE Transactions on Reliability*, 38(3):305 – 311.

(Ingongngam and Rungsawang, 2003) Ingongngam, P. and Rungsawang, A. (2003). Report on the TREC 2003 experiments using web topic-centric link analysis. In *TREC*, pages 363–367.

(JBD) JavaBDD http://javabdd.sourceforge.net

(Jeh and Widom, 2003) Jeh, G. and Widom, J. (2003). Scaling personalized web search. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA. ACM Press.

(JUN) JUNG framework http://jung.sourceforge.net

(Kao *et al.*, 2002) Kao, H.-Y., Chen, M.-S., Lin, S.-H., and Ho, J.-M. (2002). Entropy-based link analysis for mining web informative structures. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*, pages 574–581, New York, NY, USA. ACM Press.

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

(Kohlas and Haenni, 1996) Kohlas, J. and Haenni, R. (1996). Assumption-based reasoning and Probabilistic Argumentation Systems. Technical Report 96–07, Institute of Informatics, University of Fribourg. also published in "Defeasible Reasoning and Uncertainty Management Systems: Algorithms".

(Kschischang *et al.*, 2001) Kschischang, Frey, and Loeliger (2001). Factor graphs and the sum-product algorithm. *IEEETIT: IEEE Transactions on Information Theory*, 47.

(Langville and Meyer, 2004) Langville, A. N. and Meyer, C. D. (2004). Deeper inside PageRank. *Internet Mathematics*, 1(3):335–400.

(Laskey, 2005) Laskey, K. B. (2005). First-order Bayesian logic. (draft from 4/28/05).

(Milgram, 1967) Milgram, S. (1967). The small world problem. *Psychology Today*, 61.

(Newman, 2003) Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

(Page *et al.*, 1998) Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project.

(Pandurangan *et al.*, 2002) Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using PageRank to characterize Web structure. In *COCOON '02: Proceedings of the 8th Annual International Conference on Computing and Combinatorics*, pages 330–339, London, UK. Springer-Verlag.

(Pearl, 1988) Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

(Picard, 1998) Picard, J. (1998). Modeling and combining evidence provided by document relationships using Probabilistic Argumentation Systems. In *Proceedings of the ACM SIGIR'98 Conference*.

(Picard, 2000) Picard, J. (2000). *Probabilistic Argumentation Systems Applied to Information Retrieval*. PhD thesis, Neuchatel University.

(Picard and Savoy, 2003) Picard, J. and Savoy, J. (2003). Enhancing retrieval with hyperlinks: a general model based on propositional argumentation systems. *J. Am. Soc. Inf. Sci. Technol.*, 54(4):347–355.

(Poole, 2003) Poole, D. (2003). First-order probabilistic inference. In Gottlob, G. and Walsh, T., editors, *IJCAI*, pages 985–991. Morgan Kaufmann.

(Redner, 1998) Redner, S. (1998). How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4:131.

(Redner, 2004) Redner, S. (2004). Citation statistics from more than a century of Physical Review. Technical report, arxiv.org.

(Richardson and Domingos, 2001) Richardson, M. and Domingos, P. (2001). The intelligent surfer: Probabilistic combination of link and content information in pagerank. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *NIPS*, pages 1441–1448. MIT Press.

(Russell and Norvig, 2003) Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Pearson Education.

(Savoy and Picard, 1999) Savoy, J. and Picard, J. (1999). Report on the TREC-8 experiment: Searching on the Web and in distributed collections. In *TREC*.

(Savoy and Rasolofo, 2000) Savoy, J. and Rasolofo, Y. (2000). Report on the TREC-9 experiment: Link-based retrieval and distributed collections. In *TREC*.

(Shafer, 1976) Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, New Jersey.

(Shafer, 1990) Shafer, G. (1990). Perspectives on the theory and practice of belief functions. *Int. J. Approx. Reasoning*, 4(5-6):323–362.

(Upstill *et al.*, 2003) Upstill, T., Craswell, N., and Hawking, D. (2003). Predicting fame and fortune: PageRank or Indegree? In *ADCS2003*, Canberra, Australia.

(Watts and Strogatz, 1998) Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'smallworld ' networks. *Nature*, 393.

(Yedidia *et al.*, 2003) Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2003). Understanding belief propagation and its generalizations. pages 239–269.

(Zipf, 1949) Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, Cambridge.