Graph and Matrix Metrics to Analyze Ergodic Literature for Children

Eugenia-Maria Kontopoulou
CEID, University of Patras
Rio, Greece
kontopoulo@ceid.upatras.gr

Maria Predari
CEID, University of Patras
Rio, Greece
predari@ceid.upatras.gr

Thymios Kostakis
SEOKO
28, A. Papandreou st.
Halandri, Athens, Greece
kostakth@gmail.com

Efstratios Gallopoulos
CEID, University of Patras
Rio, Greece
stratis@ceid.upatras.gr

ABSTRACT

What can graph and matrix based mathematical models tell us about ergodic literature? A digraph of storylets connected by links and the corresponding adjacency matrix encoding is used to formulate some queries regarding hypertexts of this type. It is reasoned that the Google random surfer provides a useful model for the behavior of the reader of such fiction. This motivates the use of graph and Web based metrics for ranking storylets and some other tasks. A dataset, termed CHILDIF, based on printed books from three series popular with children and young adults and its characteristics are described. Two link-based metrics, `SMrank` and versions of PageRank, are described and applied on CHILDIF to rank storylets. It is shown that several characteristics of these stories can be expressed as and computed with matrix operations. An interpretation of the ranking results is provided. Results on some acyclic digraphs indicate that the rankings convey useful information regarding plot development. In conclusion, using matrix and graph theoretic techniques one can extract useful information from this type of ergodic literature that would be harder to obtain by simply reading it or by examining the underlying digraph.

Categories and Subject Descriptors

H.5.4 [**Information Systems**]: Hypertext/Hypermedia

General Terms

Algorithms, Experimentation, Human Factors, Measurement

HT'12, June 25–28, 2012, Milwaukee, Wisconsin, USA.

Keywords

hypertext, ergodic, interactive fiction, storylet, link analysis, directed graph, matrix function, stochastic matrix, path problem, ranking, PageRank, SMRank

1. INTRODUCTION

The recent issue by American Girl of the "Innerstar University books"[1], a series of printed hypertexts, at a time when the modeling and study of "connectedness" is attracting the interest of many researchers, triggered this study.

We consider some metrics based on matrix analysis and graph theory for analyzing large scale networks and study what they tell us regarding such texts. A longer term goal is to study to what extent such metrics can be used to discover interesting information, possibly uncover latent meanings, provide unexpected interpretations and hopefully assist the writer of hypertext readings cope with what is sometimes called "runaway branching".

Following Aarseth, we use the term "ergodic literature" for the genre of these books, underlying that "there is non-trival effort to traverse the text" (p.1 of [1]); cf. [24, 31, 32, 33] for information on this and the related concept of interactive fiction. In our case, when finishing reading a page of these books, one does not simply turn to the next one. Instead, the reader follows one of several possibilities, usually based on some criterion, for instance the reader's answer to one or more questions related to the story, a riddle, etc. So at the very least, the reader has to do some work ("ergon") to accomplish the page hopping necessary to find his way ("hodos") through the text. Since the way is by and large determined by the reader's decisions, these books inevitably offer plot variability.

It is fair to say that ergodic literature, especially in printed form, is not widespread and that with few notable exceptions such texts are mostly experimental and rather esoteric[2]. In

[1] `http://store.americangirl.com/agshop/html/thumbnail/id/1505/uid/814`

[2] An often cited example is Julio Cortázar's "Rayela" (Hopscotch), offering at least two possible readings. The much more recent "A Heartbreaking Work of Staggering Genius" [17] by Dave Eggers offers the reader several choices early on, while making some caustic but humoristic remarks about Interactive Fiction.

Table 1: The three collections. The first row for each entry shows the title, publisher, number of books in the series and number of books we used in this study. The second row shows years in publication and URL.

Choose Your Own Adventure (CYOA corp.) 1979-1998, `www.cyoa.com/`	>200 (6)
Innerstar University (American Girl) 2010-today, `http://web.innerstaru.com/`	9 (4)
Multiclone Tales (Kalendis pub.) 1997-2003, `http://www.kalendis.gr/`	2 (2)

fact, it has been proposed by some scholars that this state of affairs might be the downside of offering choices to the reader; see for example [16] and the discussion on page 170 of [1]. On the other hand, since the mid'70s there have been several printed collections from children pre-teens and teens, the one by American Girl being the most recent (for other examples see e.g. (e.g. [12, 34]) that are enjoying considerable success. It thus appears that young readers have been much more appreciative of the genre. Therefore, we make this literature the focus of our investigation, using books from two collections published in English and one in Greek as summarized in Table 1 (cf. Section 3.1).

Details (number of pages, possible readings, etc.) vary but their structure is similar. A plot (cf. [5]) is "woven" by putting together a "valid" sequence of pages, from starting to some terminal page[3]. Hopefully, this stimulates the children to read by engaging them in the "construction" of a plot. Following [5], we call a specific sequence of linked pages "reading". The child can read the book multiple times, from entry to some terminal page, without repeating the same reading twice. Even an adult might be under the impression that the book offers many plots to last for a long time, if not forever. This characteristic is also not lost on publishers, who frequently use it to promote this literature[4].

For the purposes of our discussion we propose the term *storylet* for the material (textual or visual) that is contained within a single page of these books (other than the usual front and back matter). This name, we hope, carries the flavor "children's tales meet computer science". Our "storylets" are a special case of Barthesian "lexias" [2] or "substories" [14]). We focus on printed hypertext books where, with the exception of front and back material, each page consists of a storylet. It is worth noting that the proper choice (content and size) of storylets is an important factor when building hypertexts [14]. In our case, we suspect that the anticipated age of readers played a role in selecting lexias to be delimited by the page boundaries, in contrast to much more aggressive linking that can be found even in early electronic hypertexts[5]. Every storylet can be one of the following: A "starting storylet", marking the beginning of a reading, an "ending storylet" marking the end of one or more readings, or a storylet of intermediate action. All storylets finish with a "branching" statement that links to other storylets, usually based on reader input, or are "endings". Sometimes, an ending storylet is exactly that; it is also possible for an ending storylet to link to the starting storylet, in case the reader wants to start over. Plot, then, is the sequence of events presented in a reading from a starting to an ending storylet. Unless specified otherwise, we consider that the first page in all readings is a starting storylet and the last one an ending storylet.

The contributions of this paper, on the subject of stories written as hypertexts modeled as directed graphs (with special emphasis on a dataset, termed CHILDIF, consisting of 12 books for children as shown in Table 1) are the following: 1) It is reasoned that there is an analogy between the actions of random surfers on the Web and readers of interactive fiction. 2) Several questions related to the characteristics of these hypertexts are shown to be expressed in terms of graph and matrix theory. 3) Because of (1), it is argued that it is sensible to consider ranking storylets. 4) A graph based metric, called `SMRank` is used to rank storylets and its properties discussed. 5) Because of (1) and the difficulty of computing (3) when graphs contain cycles, it is argued that spectral rankings like PageRank offer a practical alternative for ranking storylets. 6) The storylet rankings for all the stories in the CHILDIF dataset are computed. An interpretation of the ranking results is provided. Results on some acyclic digraphs indicate that the rankings convey useful information regarding plot development.

To the best of our knowledge, this is the first time that a systematic study of link-based analysis is conducted across a variety of literary works of interactive fiction. We note, however, that in early unpublished work Bruckman had considered the combinatorial explosion of possibilities in interactive fiction based on a graph interpretation of such stories, one of which was from the CYOA series [11].

2. DIGRAPH READINGS, SPARSITY AND SURFING

The books under study can be encoded and represented as directed graphs (digraphs)[6]. Digraphs are present in the underlying structure of several hypertext related systems, ranging from author tools (e.g. [4, 13]) to plain plot sketches by fans[7].

In the digraph model of hypertext, storylets correspond to vertices (or nodes); the starting one is a source vertex (of zero indegree), the ending ones are sinks (of zero outdegree) and the edges correspond to valid transitions between storylets. Nodes that have zero in- and outdegree (number of in- and outgoing edges are zero) play no role in the plots (they are usually visual material). Any node that has zero indegree is either a source or the result of an error. It is worth noting that the graph representation helped us in the

[3]After all, repeated readings are necessary to perceive hypertextuality, as noted in [5].

[4]From the cited literature, only [34] contains a seemingly plausible number - contained in the title - regarding the total number of possible readings. We have not had the chance to check it but it is interesting that the book also discusses, using computer science terminology some of the difficulties related to its creation.

[5]See e.g. Michael Joyce's "Afternoon, a story", `http://www.eastgate.com/catalog/Afternoon.html`, one of the first works of hypertext literature.

[6]It is worth noting that a founding member of the group Oulipo (acronym for "Ouvroir de littérature potentielle"), that has been experimenting since 1960 with novel types of writing, some resembling the ergodic genre, was the distinguished graph theorist Claude Berge. Raymond Queneau, Georges Perec and Italo Calvino were some of the well-known authors of Oulipo [7].

[7]http://www.gamebooks.org/

past to locate non-starting storylets whose nodes had zero indegree and thus were unreachable from the source; this was not what the author had in mind, and a correction was necessary.

We remind the reader (cf. [21]) that walks between two distinct vertices of a digraph are sequences of edges, where every edge and the one immediately following it share a common vertex. If the initial and last vertices in a walk are the same, the walk is closed. Trails are walks that do not repeat edges. Paths are trails without repetition of internal vertices. A cycle is a closed path of length at least 1.

Let $G = (V, E)$ be a digraph consisting of the set of vertices (nodes) V and the set of directed edges E. We assume that the digraph is simple, that is it contains no self-loops or multi-edges. For each digraph we can construct its adjacency matrix A that is a square matrix of order n, where $n = |V|$ is the number of nodes, and such that it contains 1 in position (i, j) if there is an edge from node i to node j. It has long been known that for any k, the k^{th} power of A, that is A^k, reveals in each position (i, j) the number of walks of length k between nodes i and j of the digraph; cf. [18] and [10, 29]. For all books under consideration in this study, there are only few edges leaving every node, therefore all graphs and their adjacency matrices are sparse. We denote by $\texttt{nnz} = |E|$ the total number of edges in the graph, that is the number of nonzeros in A and use $A^\top$ to denote the transpose of A.

Digraphs and sparse matrices are used to represent the connectivity and analyze the World Wide Web [9, 23], social and other networks, as well as for purposes of information retrieval; see e.g. the recent monograph [29]. Sparse matrix technology, of course, is an important topic in high performance scientific computing and can be used for the analysis of large scale networks. It is thus of interest to find out if this technology as well as tools from graph theory can be leveraged in order to learn more about the underlying books.

To this effect, we enumerate some issues, relevant to readers and storytellers, authors and publishers, in lay terms as well as in the technical language used for graphs. The list is certainly not exhaustive, there are many more interesting questions, but illustrates that one could apply the mature armory of graph and matrix theory (e.g. [20, 21, 28, 29]) to reveal the complexity of some of these questions and the means for computing exact or approximate answers.

1. How many different readings are there?
2. What is the length of each reading? Which readings are shortest/longest?
3. Is any storylet repeated in a single reading?
4. Can we rank storylets?

In this work we focus on the last question. Nonetheless, before initiating our discussion on ranking methods, let us consider for a moment some of the preceding ones. Let G be the digraph corresponding to the book under study. We assume throughout, as is the case with all books in CHILDIF, that all digraphs are simple and that there is a single source node. Regarding question (1), we need to compute how many walks are contained in G. Without additional constraints for the digraph, the walks can be many (an infinite number, if we permit cycle repetition) and counting them is hard or pointless. If G is a directed acyclic graph (DAG), however, counting walks is the same as counting paths. We will show one method shortly. Regarding (2), for general digraphs the first part is at least as hard as the previous question. However, the shortest readings amount to computing the shortest paths from the source to the sinks, a problem that can be solved with a variety of algorithms, see e.g. [19]. Finding the shortest and longest paths of a DAG with single source can also be done efficiently. Regarding question (3), it amounts to detecting cycles in the digraph, which can be accomplished, for example, by depth-first search to compute the strongly connected components of the digraph or by considering powers of the adjacency matrix, e.g. [27, 35]. Note that a nonzero diagonal element in any A^k implies the presence of a cycle. Without loss of generality in this section we assume that the nodes are numbered in such a way that the last $f \geq 1$ of them are sinks (that is ending storylets). For DAGs, this can be the result of an easy to perform relabeling of the nodes after a topological sort. Therefore, assuming that the (possibly reordered) nodes are labeled $V = \{v_1, ..., v_n\}$, then v_1 is the source (it usually is, that is the starting storylet is the first page right after the introductory material that we do not count anyway) and the last f nodes are the sinks. Then the adjacency matrix can be partitioned to have the following block form (the 0 submatrices indicate that the sinks have 0 outdegree), that is sustained when taking powers, as the right term shows:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & 0 \end{pmatrix}, \quad A^k = \begin{pmatrix} A_{11}^k & A_{11}^{k-1}A_{12} \\ 0 & 0 \end{pmatrix},$$

where A_{11}, A_{22} are both square of order $n - f$ and f respectively. The first shows the connections strictly between all nodes that are not sinks, the latter between sinks. Because $A_{11}^k = 0$ for $k \geq n - f$, the number of paths from source to each sink are readily available and denoted below by p_f, in locations $n - n_f + 1$ up to n of the first row of

$$B := \sum_{j=1}^{n-f} A^j = \begin{pmatrix} \sum_{j=1}^{n-f-1} A_{11}^j & (\sum_{j=0}^{n-f-1} A_{11}^j)A_{12} \\ 0 & 0 \end{pmatrix},$$

while the total number of paths from source to the sinks is the sum of these values. Because the DAG's adjacency matrix is nilpotent, we can write $B = (I - A)^{-1}$ and thus

$$p_f^\top = [\pi_{n-f+1}, ..., \pi_n] = e_1^\top (I - A)^{-1} \begin{pmatrix} 0 \\ I_f \end{pmatrix},$$

where I_f is the identity matrix of order f, $e_1 = [1, 0, \ldots, 0]^\top$ of order n. Therefore, the number of paths from source to the sinks is equal to $p_f^\top e^{(f)}$, where $e^{(f)}$ is the f-size vector of all 1's. Moreover

$$(I - A)^{-1} = \begin{pmatrix} (I - A_{11})^{-1} & (I - A_{11})^{-1}A_{12} \\ 0 & I_f \end{pmatrix} \qquad (1)$$

where we abuse notation slightly and do not put a subscript to the identity of similar order as A_{11}.

3. STORYLET RANKING: DAGS, SMRANK AND PAGERANK

Ranking graph nodes using matrix methods is an important topic today because of the Web and other large scale networks, though, the topic has been of interest in the social sciences much before the advent of the Internet; cf. [36]. Ranking of Web nodes is based on a complex combination of "signals", that includes the result of link-based algorithms,

such as Google's PageRank, or others [25]. Typically, this result could be represented as a nonnegative vector (sometimes stochastic[8]) with each element containing the rank of the node corresponding to that location. The question is whether ranking storylets is meaningful; and if it is, how to produce a "correct" ranking? The rest of the paper is devoted to these issues. We do this by presenting ranking methods and applying them on CHILDIF.

One way to rank nodes is by counting the in- and outdegrees of each node. These are easy to compute (in matrix terms, a simple summation of the columns and rows of the adjacency matrix provides the result, that is $A^\top e$ or Ae where e is the vector of all 1's). Since the maximum indegree for all and the maximum outdegree for all but 2 of the books in the collection was 5, this type of ranking would be too coarse to be useful. Another class of methods is based on metrics for hypertexts using distance matrices [6].

For the ergodic literature hypertexts that we consider, an alternative idea is to consider ranking schemes based on the level of storylet participation in all possible plots. So let us define a scheme in which the rank of every storylet is determined by the number of plots containing it. After normalization, we call this ranking `SMRank`.

DEFINITION 1. *Let $G = (V, E)$ be a DAG. For every $v_j \in V$ with $n = |V|$ let*

$$\tau_j := \#(\textit{paths in } G \textit{ containing } v_j).$$

Then $\texttt{SMRank} : V \to \mathbb{R}$ *is defined by*

$$\texttt{SMRank}(v_j) = \frac{\tau_j}{\sum_{j=1}^{n} \tau_j}.$$

As mentioned earlier, we assume single source digraphs. When these are DAGs, since every plot must terminate at some sink, the rank of any node can be computed by counting the total number of paths from that node to all sinks and multiplying by the total number of paths from the source to that node. If the DAG is topologically sorted making sure that all f sinks are numbered last, then $e_1^\top (I - A_{11})^{-1} e_1 = 1$ from (1) and thus $\texttt{SMRank}(v_1)$ is the sum of all the paths to the sinks. Also if $n \geq j > n - f$ (that is the node is a sink) then $\texttt{SMRank}(v_j)$ is the first element in column $j - (n - f)$ of matrix $(I - A_{11})^{-1} A_{12}$. All these results facilitate the computation of this ranking on DAGs. The following can be shown easily:

PROPOSITION 1. *Assume that the graph $G = (V, E)$ is a DAG with a single source, and that there are f sink nodes, labeled v_1 and $v_{n-f+1}, ..., v_n$ respectively, where $n = |V|$. Then it holds that for any $v_j \in V$*

$$\tau_j = e_1^\top (I - A)^{-1} e_j e_j^\top (I - A)^{-1} \begin{pmatrix} 0 \\ I_f \end{pmatrix} e^{(f)}. \qquad (2)$$

Moreover

$$\texttt{SMRank}(v_j) \leq \texttt{SMRank}(v_1).$$

Unfortunately, computing `SMRank` for an arbitrary digraph would be much more expensive (see the discussion regarding question (1) in the previous page) since we must count walks (cycles are allowed). It is reasonable to prevent traversing a cycle more than once, but still, the cost of computing the rank is prohibitive. Also, expression (2) relies on $I - A$ being invertible. This cannot be guaranteed for general digraphs (e.g. when cycles are present). Another idea is to modify the matrix by introducing a positive scaling attenuation factor say γ, such that the inverse of $I - \gamma A$ exists. Such a γ always exists for non-trivial A, by selecting any value $\gamma < 1/\rho(A)$, where $\rho(A)$ is the spectral radius of A. Then

$$(I - \gamma A)^{-1} = \sum_{j=0}^{\infty} \gamma^j A^j$$

and γ penalizes longer paths by damping their effect as powers increase. This metric was proposed early on in the field of sociometrics by Katz to rank the status of individuals in a community [22], except that he was interested in column sums of $(I - \gamma A)^{-1}$ to account for the indegrees of nodes. Thus the ranking vector was based on solving the linear system $(I - \gamma A^\top)^{-1} e$. Instead, our goal here is to rank based on the direct or indirect role of each storylet in readings, rewarding storylets that participate by providing one more step towards one or more endings (the more, the better). Notice that the attenuation factor also penalizes cycles. Their contribution is especially discounted when these are taken multiple times. In fact, even though at first sight, the penalty imposed on longer walks might seem questionable (after all we are talking about readings, not distant acquaintances), we argue that it is justified in the hypertext context. Specifically, longer stories are more likely to be stopped prematurely because the young reader rapidly loses patience or is distracted. The above metric is based on the matrix resolvent $(I - \gamma A^\top)^{-1}$. It would be natural to consider other functions as well; cf. [3, 15].

Our next goal is to propose spectral ranking and in particular a PageRank approach to rank storylets. We first describe three important modifications that are made to the graph adjacency matrix. These are standard in PageRank (cf. [25]) but we need to examine their role and justify them for the problem under study. The first is designed to prevent problems caused by "dangling nodes" (in PageRank terminology), that is sink nodes, which in our case are the ending storylets. Specifically, unless we compute `SMRank` (which is not affected by dangling nodes), we assume that from every ending storylet, the reader is offered the option to start again reading from the beginning and thus the graph is adjusted to include a link from all sink nodes to the source; the adjacency matrix is also adjusted to reflect these new links. We believe that this adjustment is reasonable and justified; it is worth noting, for example, that these "backlinks" actually exist in the text of the "Multiclone Tales".

More significantly, this modification makes the digraphs strongly connected (any node is accessible from any other node). This means the modified matrix irreducible which has important consequences as we will soon see. The second modification is to normalize the adjusted matrix to make it stochastic. This can be accomplished by dividing each row by the sum of the corresponding nodal outdegree; the resulting matrix is $D^{-1}A$, where D is the diagonal matrix of nodal outdegrees while now A is the modified matrix. Since S is stochastic, it can be considered as a transition probability matrix for a discrete-time, finite-state, stationary, irreducible Markov chain. In the sequel we also prefer to work with the transpose $S := (D^{-1}A)^\top$. Because S is column stochastic and irreducible, it has a unique positive

[8]Meaning that its elements add to 1 so that they can be interpreted as probabilities.

eigenvalue equal to 1. From Perron-Frobenius theory, the corresponding (right) eigenvector normalized to be stochastic, is unique and positive; cf. [28]. This is called the Perron vector of S and can be used to rank nodes of the graph [25] thus, in our case, the storylets of the book. Note that unlike the case of matrices representing large scale network connectivity and the Web, the matrices we are confronted with only needed a small and natural modification to guarantee irreducibility and hence the existence of a unique Perron vector. Moreover, their size is such that the vector can be computed easily using direct or iterative methods that are available in high quality numerical libraries and environnments such as MATLAB[9], our preferred package in this work. As we note in the next section, the irreducible matrices, S, resulting following the above adjustments for all the books we considered have only one eigenvalue of modulus 1 (actually real positive and equal to 1 as discussed above), therefore all matrices are primitive.

Therefore, at first sight, there appears to be no need to undertake the third proposed modification, that is to construct and use (as in PageRank) the parametric Google matrix $G(\mu) = \mu S + (1-\mu)H$, where H is the teleportation matrix and $0 < \mu < 1$ the damping coefficient that was used to guarantee that the Perron vector of $G(\mu)$ is well defined (even if S were irreducible). On the other hand, we recall that one ingenuity in PageRank was that working with G rather than S for μ strictly less than 1 also provided an interpretation of PageRank in terms of a "random surfer". In the words of Google's Brin and Page ([8]) the random surfer "is given a web page at random and keeps clicking on links, never hitting 'back' but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank." We observe that this pattern of behavior resembles that children reading this type of ergodic literature and illustrate the proposed analogy in Table 2. We thus consider the random surfer to be a suitable model for our young readers and will use the Perron vector of $G(\mu)$ to rank the storylets. It is also known that when computing PageRank, μ can be intepreted as a damping factor on distant connections, as the previously mentioned attenuation parameter by Katz; cf. [25, 36].

Another indication that this approach is justified is that the overall concept behind PageRank makes sense in our context. Specifically, it is natural to reward storylets that immediatly follow an important storylet (signifying a sequence of important events broken in storylets) but if a storylet branches out to several other ones, the rewards to storylets that follow are reduced.

It is worth noting that more detailed models could be formed by utilizing probabilistic finite automata (cf. [26]) but this is left for future investigation. Some brief comments in [30] is the closest available to the mathematical modeling of the literature we are discussing herein.

In the remainder of this paper we will apply and study the results of these rankings to determine experimentally how they work in practice. Of course, the ranking of storylets (or websites), by people or automatically (using algorithms designed by people) are inherently subjective endeavors. In evaluating the methods discussed, we need to "understand the data" well, in this case the books in CHILDIF, in order to be able to comment upon and possibly interpret the results. After all, we have an advantage over evaluating ranking results from websites, in that storylets consist of narrative elements that are part of one or more plots as opposed to website chains.

[9] `http://www.mathworks.com/products/matlab/`

3.1 The CHILDIF collection

We describe in some detail the books in the collection tabulated in Table 1.

Innerstar University This is a recent entry in the realm of ergodic children's literature. We analyzed 4 books from this collection. All of them claimed more than 20 endings in their cover. The digraphs from all the books we considered were acyclic, with a single starting storylet and several ending storylets (around 25).

Choose Your Own adventure (CYOA) The blueprint for this series was "Sugarcane Island" by E. Packard, published in 1976[10] (not included). All the books we considered were authored by R.A. Montgomery. Apparently, the series was very successful and widely translated[11]. We analyzed 6 books from this collection, 3 of which were DAGs. The number of endings in each book is shown on the cover page but no estimate of the number of stories is provided.

Multiclone Tales The last two books in our set were authored by Dr. Eugene Trivizas, one of the most prolific and beloved authors of children's literature in Greece. Trivizas, whose passion for word coinage and play is evident throughout his oeuvre, refers to these books as ΠΑΡΑπολυΜύθια (PARApolyMythia, a contrived term which could mean "too much fabling" or "polyfabling", amongst other things) while the inside cover informs us that they belong to the series "Πολύκλωνα Παραμύθια" (Multiclone Tales). Both books carry the subtitle "A magic tale with one thousand tales hidden in the same tale" and inform readers that "it is a strange and rare book that every time you read it, it tells you another story". In all respects, these two books were the most complex (none of them were DAGs) of our dataset. The smaller of the two contained two sink nodes that did not correspond to ending storylets; we did not attempt to introduce links to amend this omission. As is apparent from Fig. 3 as well as the counts in Table 5, "33 Pink Rubies" was the most complex book, providing the reader with over 220,000 possible stories[12].

We analyzed 12 books in total (a little over 1500 pages) whose titles are listed in Table 3 and created the corresponding adjacency matrices and graphs and computed several of their characteristics. We call the collection CHILDIF. In all cases we did not account for pages corresponding to nodes of zero degree (these were invariably drawings whose removal

[10] See `http://www.gamebooks.org/show_item.php?id=162` and D. Katz's `http://www.gamebooks.org/`.

[11] According to the Wikipedia entry for CYOA, more than 250 million copies were sold between 1979 and 1998.

[12] This is corroborated from anecdotal, "user experience": As one parent posted (in Greek) at `http://www.greekbooks.gr/books/pedika/pedika/ta-33-roz-rubinia.product`, "My daughter and I began this book when she was three and got bored when she was seven. Incredible imagination!".

Table 2: Reader-surfer analogy

reader	surfer
open book	turn browser on
goto starting storylet	open homepage
choose "next" storylet	click on existing link
reached ending, go to starting storylet	reached sought page, click homepage
reached ending, stop	reached sought page, stop
choose any storylet	enter URL, click preferred website

did not affect the readings). We also computed the top few eigenvalues. The second largest (in modulus) eigenvalues ranged in size from 0.88 up to 0.97, whereas the largest eigenvalue was always 1, as expected. Fig. 1 depicts the

Table 3: Book titles, abbreviated name and adjacency matrix characteristics (order n and number of nonzeros nnz) in CHILDIF. Zero degree nodes are omitted.

title	name	n	nnz
Choose Your Own Adventure			
Abominable Snowman	CYOA_AS	91	93
Journey Under the Sea	CYOA_JU	101	109
Space and Beyond	CYOA_SB	115	119
Lost Jewels of Nabooti	CYOA_LJ	110	114
Mystery of the Maya	CYOA_MM	113	116
House of Danger	CYOA_HD	91	90
Innerstar University			
Girl's Best Friend	INUN_GB	110	116
Taking the Reins	INUN_TR	103	107
Into the Spotlight	INUN_IS	112	115
Fork in the Trail	INUN_FT	110	122
Multiclone Tales			
88 Dolmadakia	MUTA_ED	154	265
33 Pink Rubies	MUTA_TP	211	386

sparse matrices INUN_GB and CYOA_SB that correspond to the two graphs shown in Fig. 1 but adjusted for the use of PageRank, that is adding links from the sinks back to the root.

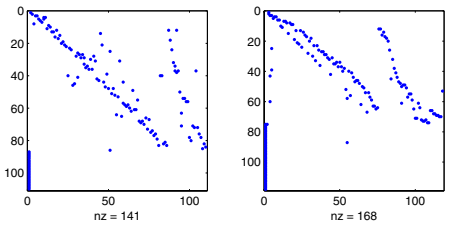

Figure 1: Matrices INUN_GB and CYOA_SB.

We next illustrate some of the books using their digraph and adjacency matrix representations. We used the visualization package GraphViz[13] and the following MATLAB toolboxes: a) Toolbox MatlabBGL[14] b) GraphViz [15] library. c) Brain Connectivity Toolbox[16]. This was used to compute lower bounds to the number of walks in graphs that were not DAGs. The graphs for some books from the collection are depicted in Fig. 2, 3 and 5.

Table 4: Books from CHILDIF that are DAGs and their characteristics. DAGness tested with MATLAB function test_dag.m from MatlabBGL.

title	storylets	ends	plots	lengths (min, max, avg)
CYOA_AS	91 (+25)	28	36	(7,20,11.8)
CYOA_MM	113 (+18)	39	106	(8,30,19.7)
CYOA_HD	91 (+17)	20	20	(9,19,14.4)
INUN_GB	110 (+11)	24	68	(8,24,17.5)
INUN_TR	103 (+19)	24	37	(10,25,16.3)
INUN_IS	112 (+11)	24	40	(7,27,17.8)
INUN_FT	110 (+11)	23	511	(7,34,25.9)

Table 5: Books from CHILDIF that are not DAG and their characteristics.

title	storylets	links	ends	plots
CYOA_JU	101 (+16)	109	42	(>202)
CYOA_SB	115 (+16)	119	43	(>98)
CYOA_LJ	110 (+21)	114	38	(>92)
MUTA_ED	154	265	41	(>1349)
MUTA_TR	211	386	53	(>220431)

4. EXPERIMENTS

In the sequel, we denote the PageRank ranking with value μ by PR(μ). As explained in Section 3, when running PR all matrices were adjusted to include links from every sink to the source. In all experiments the teleportation matrix was chosen to be $H = \frac{1}{n}ee^\top$ modeling an (impatient or bored) child that prefers to go to any other page in the book with probability $1-\mu$. We used the values $\mu = 0.85$ and 1. The latter case is simply ranking based on the Perron vector of S (thus the "bored" attitude is not captured).

Our first experiment is with "Girl's Best Friend" which is

[13] http://www.graphviz.org/

[14] Authored by D. Gleich; http://www.mathworks.com/matlabcentral/fileexchange/10922.

[15] Function graph_to_dot.m; toolbox authored by L. Peshkin; http://www.mathworks.com/matlabcentral/fileexchange/4518-matlab-graphviz-interface.

[16] Function findpaths.m; toolbox authored by O. Sporns; http://www.indiana.edu/~cortex/connectivity.html.

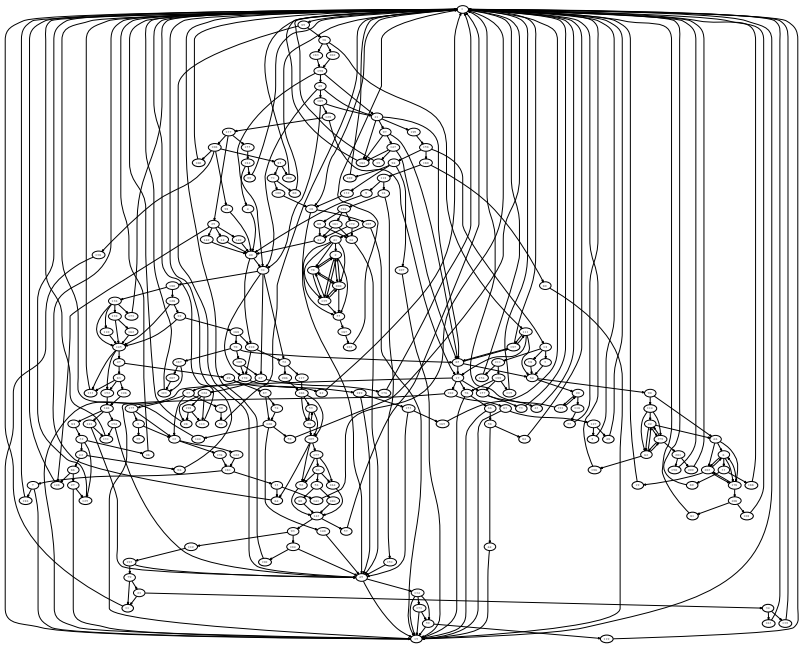

Figure 3: Graph for MUTA_TP ("33 Pink Rubies"). Sink nodes (endings) are marked with bold.

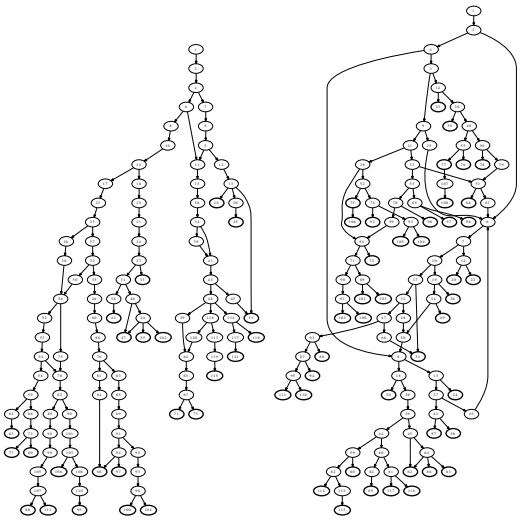

Figure 2: Digraphs for INUN_GB (DAG, left) and CYOA_JU (right) with nodes of degree 0 removed. Sink nodes (endings) are marked with bold.

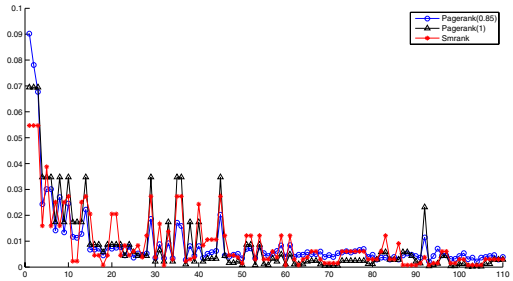

Figure 4: Rankings for all nodes of INUN_GB as computed for each ranking method. The x-axis numbers the nodes, the y-axis the ranking value.

a DAG, therefore we computed SMRank using the formulas of Section 3. For comparison we also computed the values for PR(μ) for $\mu = 0.85, 1$. The resulting rankings are shown in Fig. 4 while a more detailed view of the top results for each ranking is tabulated in Table 6. In the discussion that follows for this book the node labels are shifted down by 8 positions relative to the actual page numbers because the first storylet in the book is on page 9. Thus, node 1 corresponds to page 9, etc. As expected, the top nodes are storylets that would be part of every valid reading of any plot. Indeed, the first group of nodes may be interpreted as an initial substory, shared by all stories derived from the same novel, and for that reason it could have been written (or summarized) as a single, first, page. The second group contains node 6 (corresponding to p14). Its importance was corroborated by our "independent readers"[17]. Reading all plots, it becomes evident that node 6 takes the reader to node 11 (if one chooses for the protagonist to take Pepper the dog straight to the dog-shelter) or to node 8 (if one chooses the protagonist to walk Pepper home). This choice in some sense reveals the personality of the protagonist (that is the kid making the choice); it is therefore important, therefore the ranking algorithms were right to place the storylet higly. On the other hand, the storylets that were ranked last usually were either ending storylets or storylets directly preceding ending sto-

[17]When we refer to "independent readers" we mean friends that accepted to read the stories and mark the pages that they thought were most important, *before knowing* the results of our ranking experiments.

rylets and linked to them with a single link. Observe that in Fig. 4, there are some peaks in the PageRank values far to the right, specifically at positions 92 and 95 (the latter only for PR(0.85)). These positions correspond to nodes 51 and 68 in the graph and are both terminal nodes. The fact that their ranks are substantially higher than those of other terminal nodes (as we showed above and is validated in the figure, SMRank ranks terminal nodes last) is because they are linked directly to nodes 13 and 84, both of which lie near the root, and thus have high ranks that pump up the PageRank of their immediate descendants. Such "link jumps" exist in several of the books and manifested themselves with a similar pattern of peaks in the ranking plots.

Table 6: Top ranked results for INUN_GB. Actual page numbers have been shifted by 8 so that the initial book page of 9 is listed as 1.

SmRank		Pagerank			
		$\mu = 0.85$		$\mu = 1$	
value	node	value	node	rank	node
0.0547	4	0.0903	1	0.0695	4
	2	0.0781	2		1
	1	0.0677	4		2
0.0388	6	0.0302	6	0.0347	43
0.0274	56		7		41
	43	0.0270	9		11
	41	0.0245	11		5
	34	0.0243	5		56
	15	0.0222	15		6
	11	0.0202	56		15
0.0251	14	0.0186	34		7
	10	0.0171	41		9
	8	0.0159	43		34

We next consider "Mystery of the Maya" (CYOA_MM), also a DAG; cf. Fig. 5. As before, the first few pages ($1 \to 2 \to 3 \to 5 \to 6$) are supposed to be read in sequence. These introduce some background to the story ("a friend of the hero has disappeared while on assignment in Mexico and the hero decides to search for him"). Not surprisingly, all ranking algorithms identify these pages as being the most important. Also, sinks were consistently ranked very low. The groupings in SMRank and PR(1) for the top nodes are similar, while PR(0.85) appears to provide a more refined view, placing the source above the others. P6 contains the first decision for the reader. Moreover, after reading the book it became clear that p6 was quite special because it divides the book into two "conceptual directions". These directions are further divided in the sequel as in the concept map shown in Fig. 6. After reading all the plots in the book, three storylets (on pages 6, 12 and 38) that were judged by the readers to be important, were also assigned a high SMRank value. From the readings, it became clear that at those nodes there were "conceptual" directions that opened up. It is also worth noting that p12 and p38 are at the same level, but have very different ranks. This can be exlained by the fact that p12 leads to 24 endings whereas p38 to only 10. We show the ranks for all the nodes in Fig. 7.

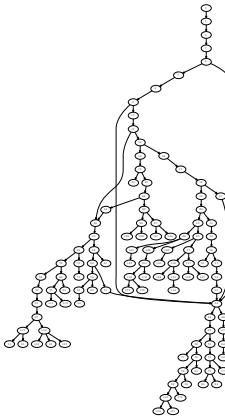

Figure 5: Graph for CYOA_MM (DAG). Nodes of zero degree are not depicted. Sink nodes (endings) are marked with bold.

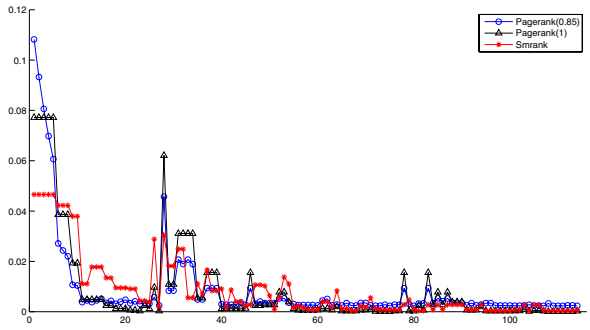

Figure 7: Rankings for all nodes of CYOA_MM.

Table 7: Top ranked results for CYOA_MM.

SmRank		Pagerank			
		$\mu = 0.85$		$\mu = 1$	
rank	node	rank	node	rank	node
0.0466	6	0.1082	1	0.0772	1
	5	0.0933	2		5
	3	0.0806	3		2
	2	0.0698	5		6
	1	0.0607	6		3
0.0422	9	0.0457	38	0.0621	38
	8	0.0271	7	0.0386	7
	7	0.0244	8		8
0.0379	12	0.0220	9		9
	11	0.0207	44	0.0310	46
0.0304	38		46		44
0.0288	34	0.0189	45		45
0.0249	45		47		47
	44	0.0107	11	0.0193	12
0.0182	43	0.0104	12		11

Another observation is that SMRank and PR(1) gave similar orderings in all DAG books under evaluation.

We next consider one book that is not a DAG, specifically CYOA_JU, and the ranking obtained by PageRank. Results are shown in Fig. 8 and Table 8. It can be seen from the

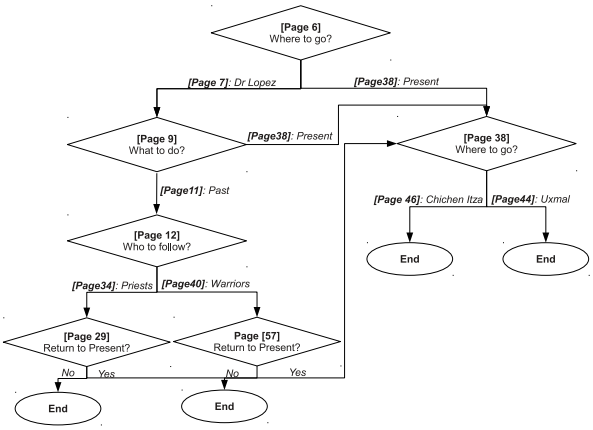

Figure 6: Concept map for CYOA_MM. Each box represents several storylets and possible plots (endings included), but sharing the marked concept.

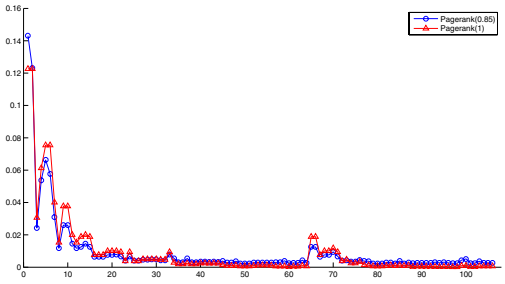

Figure 8: Rankings for all nodes of CYOA_JU.

table that a significant number of storylets that are ranked as important belong to a cycle (pages 6, 8, 7,10,13 etc). To see why this is so, we examine the rankings at the nodes where cycles begin. Consider for instance the cycle $6 \to 7 \to 10 \to 17 \to \cdots \to 27 \to 43 \to 6$. Pages 6 and 8 can be viewed as entry nodes to the cycle since any path entering the cycle has to include them. PageRank recognizes this and assigns these pages relatively higher values than the rest of the cycle nodes. For the same reason, if the entry nodes had low rank, then the rank of the remaining nodes in the cycle would be even lower. There is also no difference for the tabulated nodes between the rankings computed with `PR(1)` and `PR(0.85)`.

Finally, it is interesting to note that none of the books had more than one starting storylet. This would have made the rankings more interesting as there would be more candidates for the top position (currently occupied by the single source).

5. CONCLUSIONS

From the above discussion and experiments it appears that the `SMRank` (for DAGs) and PageRank can be used to obtain interpretable and useful information from the CHILDIF collection. This opens the way for further analysis of hypertexts of this kind using graph and matrix tools developed

Table 8: Top ranked results for CYOA_JU.

Pagerank			
$\mu = 0.85$		$\mu = 1$	
rank	node	rank	node
0.1431	1	0.1226	1
0.1231	2		2
0.0663	6	0.0754	7
0.0578	7		6
0.0537	4	0.0613	4
0.0310	8	0.0401	8
0.0260	12	0.0377	10
	10		12
0.0242	3	0.0307	3
0.0146	18	0.0200	13
	13		18
0.0125	20	0.0189	22
	19		20
	17		17
	22		19

for link based analysis, including recent methods developed for digraphs; see e.g. [3]. There are also interesting educational avenues, such as the use of this "literary framework" to introduce modern computer science and mathematical concepts for data analysis. A further challenge would be to apply matrix and graph tools online, on ergodic literature that cannot be readily leafed through[18]. The bigger quest, of course, is to incorporate such methods into tools that can assist readers, authors, and even publishers of ergodic literature, printed or digital.

6. ACKNOWLEDGMENTS

We are grateful to: Markus Strohmaier and the reviewers for considering our work and for constructive comments;

[18] As in the interactive fiction "Choice Of" series; cf. `http://www.choiceofgames.com/`.

Michele Benzi for discussions on the topic of matrix functions for network analysis and Oulipo; Giorgos Kollias for helpful discussions and advice on tools for graph manipulation; Christos Zaroliagis and Panagiotis Mihail for discussions on graph algorithms; Aristoula Georgiadou for advice on philological matters. We also thank Eugenios Trivizas (who has been Professor of Criminology at the Department of Sociology of University of Reading, UK, since 1978) for his comments on a very early version of this work presented at a local student workshop in Patras[19]. The current paper represents a major advancement over that work. We also thank our "independent readers team", Ioanna Gazi and Angeliki Rapti. Our discovery of children's hypertext literature in Greece was the result of some serendipity having to do with parenting and storytelling. So it is appropriate to also thank Anabella for causing it!

7. REFERENCES

[1] E. Aarseth. *Ergodic Literature*. The Johns Hopkins University Press., Baltimore, MD, 1997.

[2] R. Barthes. *S/Z: Essais*. Seuil, Paris, 1970.

[3] M. Benzi, E. Estrada, and C. Klymko. Ranking hubs and authorities using matrix functions. Technical Report Math/CS TR-2012-003, Emory University, 2012.

[4] M. Bernstein. Storyspace 1. In *Proc. 13th ACM Conf. Hypertext and Hypermedia*, HT'02, pages 172–181, New York, NY, USA, 2002. ACM.

[5] M. Bernstein. On hypertext narrative. In *Proc. 20th ACM Conf. Hypertext and Hypermedia*, HT'09, pages 5–14, New York, NY, USA, 2009. ACM.

[6] R. A. Botafogo, E. Rivlin, and B. Shneiderman. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Trans. Inf. Syst.*, 10:142–180, April 1992.

[7] D. Bouyssou, D. de Werra, and O. Hudry. Claude Berge and the "Oulipo". *EURO Newsletter*, (6), 2006.

[8] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 33:107–117, 1998.

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Comput. Netw.*, 33(1-6):309–320, 2000.

[10] R. Brualdi and D. Cvetkovic. *A Combinatorial Approach to Matrix Theory and its Applications*. Chapman & Hall/CRC, 2009.

[11] A. Bruckman. The combinatorics of storytelling: Mystery train interactive. Unpublished paper, MIT Media Lab in `http://www.cc.gatech.edu/~asb/papers/misc/combinatorics-bruckman-90.pdf`, 1990.

[12] K. Casey. *The Runaway Game*. May Davenport Publishers, Los Altos Hills, CA, 2001.

[13] M. P. Consens and A. O. Mendelzon. Expressing structural hypertext queries in Graphlog. In *Proc. 2nd Annual ACM Conf. on Hypertext*, HT'89, pages 269–292, New York, NY, USA, 1989. ACM.

[14] C. Crawford. *On interactive storytelling*. New Riders Games, Berkeley, CA, 2005.

[15] J. Crofts, E. Estrada, D. Higham, and A. Taylor. Mapping directed networks. *Electronic Transactions on Numerical Analysis*, 37:337–350, 2010.

[16] A. Doxiadis. The mathematical logic of narrative. In M. Manaresi, editor, *Matematica e cultura in Europa*, pages 171–181. Springer, Milan, 2005.

[17] D. Eggers. *A Heartbreaking Work of Staggering Genius*. Simon & Schuster, New York, 2000.

[18] L. Festinger. The analysis of sociograms using matrix algebra. *Human Relations*, 2:153 – 158, 1949.

[19] M. Fredman and R. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *J. ACM*, 34(3):596–615, July 1987.

[20] I. Gessel and R. Stanley. Algebraic enumeration. In R.L. Graham et al., editor, *Handbook of Combinatorics*, volume 2, pages 1021–1061. Elsevier, 1995.

[21] J. Gross and J. Yellen, editors. *Handbook of Graph Theory*. CRC Press, 2004.

[22] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43, 1953.

[23] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1999.

[24] G. Landow. *Hypertext 3.0*. Johns Hopkins University Press, 2006.

[25] A. Langville and C. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton Univ. Press, 2006.

[26] M. Levene and G. Loizou. Web interaction and the navigation problem in hypertext. In *Encyclopedia of Microcomputers*, volume 28 (suppl. 7), pages 381–398. Marcel Dekker, New York, 2002.

[27] P. Mateti and N. Deo. On algorithms for enumerating all circuits of a graph. *SIAM Journal on Computing*, 5(1):90–99, 1976.

[28] C. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 2001.

[29] R. Mihalcea and D. Radev. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge University Press, 2011.

[30] N. Montfort. Cybertext killed the hypertext star: The hypertext murder case. *Electronic Book Rev.*, 2000.

[31] N. Montfort. *Twisty Little Passages*. MIT Press, 2003.

[32] A. Saemmer. Littératures numériques: tendances, perspectives, outils d'analyse. *Études Françaises*, 43:111–131, 2007.

[33] S. Schreibman, R. Siemens, and J. Unsworth, editors. *A Companion to Digital Humanities*. Blackwell, Oxford, 2004.

[34] J. Shiga. *Meanwhile: Pick Any Path. 3,856 Story Possibilities*. Amulet Books, 2010.

[35] R. Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[36] S. Vigna. Spectral ranking. *CoRR*, abs/0912.0238, 2009.

[19] T. Kostakis and E. Gallopoulos, "The 88 dolmadakia eigenvector: Link analysis and linear algebra in children's books", Eureka Conf. presentation, Patras, 2007.