Finding Related Pages Using Green Measures: An Illustration with Wikipedia

Yann Ollivier
CNRS & École normale supérieure de Lyon
46, allée d'Italie
69364 Lyon Cedex 07, France
yann.ollivier@normalesup.org

Pierre Senellart
INRIA Futurs & Université Paris XI
4 rue Jacques Monod
91893 Orsay Cedex, France
pierre@senellart.com

ABSTRACT

We introduce a new method for finding nodes semantically related to a given node in a hyperlinked graph: the Green method, based on a classical Markov chain tool. It is general, adjustment-free and easy to implement. We test it in the case of the hyperlink structure of the English version of Wikipedia, the on-line encyclopedia. We present an extensive comparative study of the performance of our method compared to several other classical methods in the case of Wikipedia. The Green method is found to have both the best average results and the best robustness.

1. INTRODUCTION

The use of tools relying on graph structures for extracting semantic information in a hyperlinked environment [11] has had vast success, which led among other things to a revolution in the search technology used on the World Wide Web [2, 5]. In the same spirit, we present in this paper a novel application of a classical tool from Markov chain theory, the Green measures, to the extraction of semantically related nodes in a directed graph. Such a technique can help a user find additional content closely related to a node i and thus guide her in the exploration of a graph. Google [5] and Google Scholar [6] both allow users to search for similar nodes, respectively in the Web graph and in the graph of scientific publications. This can be all the more useful in the case of the graph of an on-line encyclopedia like Wikipedia [18], where articles are seen as nodes of the graph and hyperlinks as edges between nodes: users are often interested in looking for articles on related topics, for instance to deepen their understanding of some concept. Other interests of an automatic method for finding related articles can be for instance to add missing links between articles [1] and thereby to enrich the content of the encyclopedia. Our proposed method could be intuitively described as a PageRank [2] computation that continuously pours mass at node i (and uniformly removes mass everywhere else to ensure convergence).

In order to be able to have a somewhat objective measure of the performance of the Green method, we present in detail six different algorithms for extracting related pages in a graph, either based on Green measures or on other, more classical approaches. These algorithms have been implemented, and tested on the graph of the English version of Wikipedia; though, to preserve generality of the approach, we did not implement any Wikipedia-specific tricks to enhance performance. A user study has been performed which allows us to evaluate and compare each of these techniques.

Our contributions are thus: (i) a novel use of Green measures to extract semantic information in a graph (ii) an extensive comparative study of the performance of different methods for finding related articles in the Wikipedia context. Note that we implemented "pure" versions of the methods: it is certainly easy to devise Wikipedia-specific enhancements to the methods, but we refused to do so in order to keep the comparison general. Even so, performance of the Green method was very satisfying.

In Section 2, we briefly review basics of Markov chain theory, introduce Green measures, and give intuitive interpretations of them. We then present in Section 3 different methods for extracting related nodes in a graph, based on Green measures, on PageRank, and on classical Text Mining approaches. The context, methodology, and results of the experiment we carried out to compare these methods are described in Section 4. Finally, related work is presented in Section 5, and perspectives for extension and improvement of the Green method are discussed in Section 6.

Additional data about the content presented here (including source code and full evaluation results) is available on the companion website for this paper [14].

2. GREEN MEASURES

We collect here some notation and standard material about Markov chains. Informally, a Markov chain is defined by specifying a state space X and, for each $i \in X$, transition probabilities p_{ij} indicating the probability that the next state is j knowing that the current state is i.

Any graph can be viewed as a Markov chain as follows: the state space is the set of nodes, and knowing that the current state is i, the next state is chosen uniformly among the nodes j with an edge from i to j. This remark is very important since it allows to view any hyperlinked environment as a Markov chain and to use and/or adapt Markov chain techniques. The main example is of course the PageRank algorithm [2].

We now collect some standard facts and notation about Markov chains. We refer to [13], or to [8] for a more algorithmic viewpoint.

2.1 Notation for Markov Chains

Let (p_{ij}) be the transition probabilities of a Markov chain on a finite set X. That is, each p_{ij} is a non-negative number representing the probability to jump from node $i \in X$ to node $j \in X$; in particular, for each i we have $\sum_j p_{ij} = 1$. That is, the p_{ij}'s form a stochastic matrix.

For example, if X is given as a directed graph, we can define the *simple random walk on* X by setting $p_{ij} = 0$ if there is no edge from i to j, and $p_{ij} = 1/d_i$ if there is an edge from i to j, where d_i is the number of edges originating from i (if multiple edges are allowed, this definition can be adapted accordingly).

A row vector $\mu = (\mu_i) : X \to \mathbb{R}$ indexed by X will be called a *measure on* X (negative values are allowed); the value μ_i will be called the *mass* of i. The *(total) mass* of μ is $\sum \mu_i$. If moreover $\mu_i \geqslant 0$ and $\sum \mu_i = 1$, the measure will be called a *probability measure.*

We define the *forward propagation operator* M as follows: for any measure $\mu = (\mu_i)$ on X, the measure μM is defined by

$$(\mu M)_j := \sum_i \mu_i p_{ij}$$

that is, each node i sends a part p_{ij} of its mass to node j. This corresponds to multiplication on the right by the matrix $M = (p_{ij})$, hence the notation μM. Note that forward propagation preserves the total mass $\sum \mu_i$.

Henceforth, we suppose, in a standard manner, that the Markov chain is irreducible and aperiodic [13, 8]. For the simple random walk on a graph, it amounts to the graph being strongly connected and the greatest common divisor of the lengths of all cycles being equal to 1.

Under these assumptions, and since the state space is finite, it is well-known that the Markov chain has a unique invariant probability measure ν, the *equilibrium measure*: that is, a unique measure ν with $\nu M = \nu$ and $\sum \nu_i = 1$. Moreover, for any measure μ such that $\sum \mu_i = 1$, the iterates μM^n converge to ν as $n \to \infty$. More precisely, the matrices M^n converge exponentially fast to a matrix M^∞, which is of rank 1 and satisfies $M^\infty_{ij} = \nu_j$ for all i. The equilibrium measure ν can be thought of as a PageRank without random jumps on X (see [2]).

2.2 Definition and Interpretation of Green Measures

Green functions were historically introduced in electrostatic theory as a means of computing the potential created by a charge distribution; they have later been used in a variety of problems from analysis [4]. Since then, a deep analogy has been recognized between electrostatic potential theory and Markov chains [10], which allows to define discrete analogues of the Green functions (basically by replacing the continuous Laplacian with the discrete Laplacian on a graph). The Green measure centered at i, as defined below, can really be thought of as the electric potential created on X by a unit charge placed at i.

The *Green matrix* (or *potential kernel*, or *fundamental matrix*) of a finite Markov chain is defined by

$$G := \sum_{t=0}^{\infty} (M^t - M^\infty)$$

where M^t is the t-th power of the matrix $M = (p_{ij})$, corresponding to t steps of the random walk. Since the M^t converge exponentially fast to M^∞, the series converges.

Now, for $i \in X$, let us define G_i, the *Green measure centered at* i, as the i-th line of the Green matrix G.

Let δ_i be the Dirac measure centered at i, that is, $\delta_{ij} := 1$ if $j = i$ and 0 otherwise. We have by definition $G_i = \delta_i G$. More explicitly, using that $M^\infty_{ij} = \nu_j$, we get

$$G_{ij} = \sum_{t=0}^{\infty} \left((\delta_i M^t)_j - \nu_j \right)$$

where $(\delta_i M^t)_j$ is of course the probability that the random walk starting at i is at j at time t. Since $\delta_i M^t$ is a probability measure and ν is as well, we see that for each i, G_i is a measure of total mass 0.

We present two natural interpretations of the Green measures (in addition to electric potential):

PageRank with source at i: The sum

$$G_i = \sum_{t=0}^{\infty} (\delta_i - \nu) M^t$$

is a fixed point of the operator

$$\mu \mapsto \mu M + (\delta_i - \nu)$$

This fixed point is thus the equilibrium measure of a random walk with a source term δ_i which constantly pours a mass 1 at i, and a uniform sink term $-\nu$ (to preserve total mass).

This shows how Green measures can be computed in practice: Start with the row vector $\mu = \delta_i - \nu$ and iterate $\mu \mapsto \mu M + (\delta_i - \nu)$ some number of times. This allows to compute the Green measure centered at i without computing the whole Green matrix.

Time spent at a node: Since the equilibrium measure is ν, the average time spent at any node $j \in X$ by the random walk between steps 0 and t behaves, in the long run, like $(t+1)\nu_j$, whatever the starting node was. Knowing the starting node i leads to a correction term, which is precisely the Green measure centered at i. More precisely:

$$G_{ij} = \lim_{t \to \infty} \left(T_{ij}(t) - (t+1)\nu_j \right)$$

where $T_{ij}(t)$ is the average number of times the random walk starting at i hits node j between steps 0 and t (included).

3. DESCRIPTION OF THE METHODS

The goal of each method is, given a node i in a graph (or in a Markov chain), to output an ordered list of nodes which are "most related" to i in some sense. All methods used here rely on scoring: given i, every node j is attributed a score $S^i(j)$. We then output the n nodes with highest score.

Here the number n was arbitrarily set to 20 for each method, as we could not devise a natural universal way to define a threshold.

We now proceed to the definition of the scores $S^i(j)$ for the five methods included in the evaluation: **Green**, **SymGreen**, **PageRankOfLinks**, **Cosine** and **Cocitations**. We also describe some variants of these, which were implemented but not included in the evaluation in order not to ask too much from our human testers.

3.1 Two Green-Based Methods

3.1.1 Green

The **Green** method relies directly on the Green measures described above. When looking for nodes similar to node i, compute the Green measure G_i centered at i. Now for each j, the value G_{ij} indicates how much node j is related to node i and can be taken as the score $S^i(j)$.

This score leads to satisfying results. However, nodes j with higher values of the equilibrium measure ν_j were slightly overrepresented. We found that performance was somewhat improved if an additional term favoring uncommon nodes j (as measured by the value ν_j of the equilibrium measure at j) is introduced. Namely we set

$$\boxed{S^i(j) := G_{ij}\log(1/\nu_j)}$$

The logarithmic term comes from information theory. Indeed, $\log(1/\nu_j)$ is the quantity of information brought by the event "the random walk currently lies at node j", knowing that its prior probability is ν_j. Therefore, the quantity $G_{ij}\log(1/\nu_j)$ measures both how close to i and how informative node j is.

This logarithmic term is very similar in spirit to the one appearing in the tf-idf formula used for **Cosine**, as described in Section 3.3.1.

3.1.2 SymGreen

Since it mainly consists in following the Markov chain flow starting at node i, **Green** might miss nodes that point to i but are not pointed by i, nodes which could be worth considering. The workaround is to somehow "symmetrize" the Markov chain as follows: Given any Markov chain (p_{ij}) with stationary measure $\nu = (\nu_i)$, the *symmetrized Markov chain* is defined by

$$\tilde{p}_{ij} = \frac{1}{2}\left(p_{ij} + p_{ji}\frac{\nu_j}{\nu_i}\right)$$

This new Markov chain still has ν as its equilibrium measure, and moreover it enjoys the stronger property of *detailed balance* i.e. $\nu_i\tilde{p}_{ij} = \nu_j\tilde{p}_{ji}$, meaning that at equilibrium not only the mass of nodes are stabilized, but the mass flow through each edge is balanced.

This amounts to, at each step, tossing a coin and following the origin Markov chain either forward or backward (where the backward probabilities are given by $p_{ji}\nu_j/\nu_i$). For the simple random walk on a regular graph, this corresponds to forgetting the edge orientations.

The Green measures $\tilde{G}_i$ for this new Markov chain $(\tilde{p}_{ij})$ can be defined in the same way, and as above the scores are given by

$$\boxed{S^i(j) := \tilde{G}_{ij}\ \log(1/\nu_j)}$$

3.2 PageRank-Based Methods

Arguably the most naive method for finding nodes related to a given node i is to look at nodes with high PageRank index in a neighborhood of i. Similar techniques are extensively used for finding *related pages* on the World Wide Web [11, 3] (see Section 5). Here by PageRank we mean the equilibrium measure of the random walk, that is, we discard random jumps (we set Google's *damping factor* to 1). Indeed, random jumps tend to spread the equilibrium measure more uniformly on a graph, whereas the goal here is to focus around a given node.

We describe two ways of using the equilibrium measure to identify nodes related to a given node.

3.2.1 PageRankOfLinks

The first method that springs to mind for identifying nodes related to i is to take the nodes pointed by i and output them according to their PageRank.

Namely, let ν be the equilibrium measure of the random walk on the graph (or of the Markov chain). Let i be a node. The score of node j in the **PageRankOfLinks** method is defined by

$$\boxed{S^i(j) := \begin{cases} \nu_j & \text{if} \quad p_{ij} > 0 \\ 0 & \text{if} \quad p_{ij} = 0 \end{cases}}$$

The only advantage of this method is that it is very simple. However, as can clearly be seen on the results below, as soon as node i has a significant number of neighbors, it is rather unspecific and tends to output the same nodes whatever the initial node i was.

3.2.2 LocalPageRank

Another PageRank-based method was implemented. It consists in, first, building a restricted graph centered at node i, and then computing the equilibrium measure on this subgraph. The method outputs nodes of this subgraph, ordered according to this "local PageRank".

More precisely, let $i \in X$ be a node. The *neighborhood* N_i of i is defined as the set of nodes $j \in X$ which either are pointed by i, or point to i, or are pointed by a node pointing to i (backward-forward siblings), or point to a node also pointed by i (forward-backward siblings). Also, conventionally we decide that $i \in N_i$.

The subgraph N_i might not be strongly connected, so instead, let N'_i be the strongly connected component of N_i containing i. (Otherwise, we will not be able to define an equilibrium measure.)

This new graph N'_i is a subgraph of the original graph. The random walk on N'_i has an equilibrium measure $\nu^{N'_i}$, which can be thought of as a local PageRank around i.

Now the method consists in outputting the nodes in N'_i, ordered according to the value of $\nu^{N'_i}$:

$$\boxed{S^i(j) := \begin{cases} \nu_j^{N'_i} & \text{if} \quad j \in N'_i \\ 0 & \text{if} \quad j \notin N'_i \end{cases}}$$

This method has an important flaw: As soon as the graph is highly connected, as is the case with Wikipedia, the neighborhood N'_i comprises a significant portion of the original graph. In such a case, the equilibrium measure $\nu^{N'_i}$ is very close to the global equilibrium measure ν, and so the results are not at all specific to i. In the example of Wikipedia, for typical nodes i, the subgraph N'_i comprises one third to one half of the whole graph. In particular all nodes with high global PageRank belong to it, so these nodes appear first in the result though not particularly related to i.

The problem cannot be solved by taking smaller neighborhoods: N'_i is included in the 2-neighborhood of i. It is hard to discard more nodes: nodes pointed by nodes pointing to i (backward-forward siblings) as well as nodes pointing to nodes pointed by i (forward-backward siblings) really are natural candidates for nodes related to i, and precisely are the ones considered by methods such as COSINE and COCITATIONS. This problem is probably intrinsic to any highly connected graph.

Due to its extremely poor results, this method was not included in the test. For example, on *Pierre de Fermat* the first 10 results in the output are *France*, *United States*, *United Kingdom*, *Germany*, *2005*, *2006*, *World War II*, *Italy*, *Europe* and *England*, showing no specific relationship to the base article but virtually identical to the global PageRank values. (Compare the other methods' results on Table 2.)

3.3 Text Mining-Inspired Methods

Standard text mining methods can be readily applied in a graph/Markov chain setting, provided one is able to define the "content" of a node i. It is natural to interpret the set of nodes pointed by i as the content of i, and moreover the transition probabilities p_{ij} can be thought of as the frequency of occurrence of j in i. One can then use standard document manipulation and comparison methods.

We tested two such methods: a cosine method using a tf-idf weight, and a cocitation index method.

Let us insist that these methods were applied to the *hyperlink graph structure* of Wikipedia, ignoring the textual content (except of course in order to build the graph). The strength of all methods presented in this paper is precisely that they can be applied when only the graph structure is available. We think that other text mining methods may be extended to work in such a graph setting.

3.3.1 COSINE *using tf-idf weight*

Cosine computations first use some transformation to represent each node/document in the collection by a vector in $\mathbb{R}^n$ for some fixed n. The proximity of two such vectors can then be measured by their cosine as ordinary vectors in $\mathbb{R}^n$ (or their angle, which amounts to the same as far as ordering is concerned).

One very usual such vector representation for documents is given by the *term frequency/inverse document frequency (tf-idf) weight* [15]. In our setting, it is adapted as follows.

Given a Markov chain defined by (p_{ij}) on a set of N elements (e.g. the random walk on a graph), for each node i the tf-idf vector x^i associated to i is an N-dimensional vector defined by

$$(x^i)_j := p_{ij} \log\left(N/d_j\right)$$

where d_j is the number of nodes pointing to j.

COSINE is then very simple: when looking for nodes related to node i, the score of node j is defined by

$$\boxed{S^i(j) := \cos(x^i, x^j)}$$

where x^i and x^j are seen as vectors in $\mathbb{R}^N$. Here of course $\cos(x, y)$ is computed via the usual formula

$$\cos(x, y) = \frac{\sum x_k y_k}{\sqrt{\sum x_k^2}\sqrt{\sum y_k^2}}$$

We also implemented a variant where $\log(N/d_j)$ was replaced with $\log(1/\nu_j)$ with ν_j the equilibrium measure of the random walk (which is more natural in our setting), and a variant where the log term was simply removed. The results observed were not significantly different, so these methods were not included in the evaluation.

3.3.2 COCITATIONS

A standard straightforward method to evaluate document similarity is the cocitation index: two documents are similar if there are many documents pointing to both of them. This method, which originated in bibliometrics, is well-known and widely used for similar problems, see for instance [3] in the context of the Web graph.

In our context this simply reads as follows. When looking for nodes similar to a node i, the score of node j is given by

$$\boxed{S^i(j) := \#\left\{k,\, p_{ki} > 0 \text{ and } p_{kj} > 0\right\}}$$

Sometimes this method tends to favor nodes that have the same "type" as i rather than nodes semantically related to i but with a different nature. For example, when asked for pages related to *1989* (the year) in Wikipedia, the output is *1990*, *1991*... For the base article *Pierre de Fermat*, interestingly, it outputs several other significant mathematicians.

4. EXPERIMENTAL RESULTS

In this section, we describe the experiments carried out to evaluate the performance of the methods presented in Section 3, on the graph of the English version of the on-line collaborative encyclopedia Wikipedia [18].

4.1 Graph Extraction

A dump of the Wikipedia databases can be downloaded from `http://download.wikimedia.org/` as compressed XML files. We processed a September 25th, 2006 dump of the English Wikipedia in order to build the corresponding directed graph. A few precautions have to be taken:

Wiki formatting. Wikipedia articles are written using a Wiki syntax, which has to be parsed in order to remove comments and extract links and template invocations.

Normalization of article titles. For correct identification of articles pointed by links, a few Wikipedia conventions have to be followed, including capitalization of every first letter of an article title, and equivalence of underscore and space characters in titles.

Multiple links. If multiple links to an article are present in another article, this multiplicity is preserved as a weight of the corresponding edge of the graph.

Redirections. A large number of Wikipedia entries are just used as redirections to other articles. For instance, *Gödel* is a redirection to *Kurt Gödel*; *NP-Complete* is a redirection to *NP-complete*. Such redirections are not real entries. Links pointing to such redirections have been resolved.

Templates. *Templates* are used in Wikipedia to represent similar-looking content in different articles. For instance, each article about a country uses the *Infobox Country* template to display in a generic format a number of different items of information, such as the area of the country, its currency or its official language(s). Other templates are used in a more general context, such as the *Details* template, which shows a line *For more detail on this topic, see ...* We first tried building the graph by adding all links appearing in the expansion of templates; this had the tendency, however, to "pollute" the set of links of articles by adding a large number of mildly relevant links (for example, the *Infobox Country* template adds links to such general articles as *Population density*, *Time zone*...). We finally decided to expand only the generic and most used templates *Main*, *See also*, *Further*, *Details* and their variants, otherwise using only explicit links from the main text to build the graph.

Categories. *Categories* are special entries on Wikipedia which are used to group articles according to semantic categories; there is for instance a *Living people* category to which every article about a living person should belong. Categories may also contain subcategories, which allows for hierarchical structures. We chose to consider each category as a standalone node, which is consistent with the way categories are presented on the Web interface of Wikipedia: If an article a belongs to a category c, there are links both from a to c and from c to a. This tends to improve strong connectivity of the graph.

An implementation of this graph extraction procedure, in the form of a collection of C++ and Perl programs, is available from [14].

The resulting graph has $1,606,896$ nodes and $38,896,462$ edges; there are $73,860$ different strongly connected components, the largest one of which includes $1,531,989$ nodes, that is, 95% of the nodes of the graph (so that the remaining strongly connected components contain 1.01 node on average). We will restrict ourselves to this strongly connected subgraph, in order to ensure the applicability of the theoretical results presented in Section 2, and, in particular, convergence of computation of the equilibrium measure and Green measures. The other required condition, aperiodicity, is trivially guaranteed by the fact that some articles contain self-links. More detailed link analyses of the Wikipedia graph can be found in [17, 9, 19].

4.2 Implementation

The algorithms presented in Section 3 have been implemented in C++; the corresponding source code is freely available from [14]. Implementation of the methods is mostly straightforward, but here are a few caveats:

- Because of the large size of the Wikipedia graph (the size of the data structure required to store the weighted sparse adjacency matrix is about 1GB), special care has to be taken for memory handling; a large sparse graph library, relying on memory-mapped files, has been developed for this purpose.

- Most methods require prior knowledge of the equilibrium measure for the graph. The latter is computed once with very high accuracy, in order not to compromise the precision of the methods using it; we found that, after 100 iterations, the total variation distance (computed as $\frac{1}{2}\sum_i |(\nu M)_i - \nu_i|$) is about 10^{-5}.

- Storage of the full Green matrix is impractical, as it has $1,606,896 \times 1,606,896$ entries, none of which are a priori equal to 0. Therefore, it is necessary to only compute the Green measure G_i for a given input article i. For this, we use the characterization of Green measures as fixed point of the operator $\mu \mapsto \mu M + (\delta_i - \nu)$ (keeping the notation of Section 2.2). We start with the row vector $\mu = \delta_i - \nu$; then, 5 iterations of the operator are enough to ensure good convergence as far as the top values are concerned.

The computation time for **Green** is less than 10s on a 3GHz desktop PC; that of **SymGreen** is typically between 15s and 30s, whereas **Cosine** computation time ranges from one second on very short articles to more than three minutes on large articles. **LocalPageRank** computation takes from one second to more than one minute depending on article size. **Cocitations** is fast, taking at most three seconds on large articles, whereas **PageRankOfLinks** is virtually instantaneous on the same computer.

4.3 Evaluation Methodology

In order to evaluate the relative performance of the different methods presented in Section 3, we carried out a blind evaluation of their results on 7 different articles, chosen for their diversity:

- *Clique (graph theory)*: a very short, technical article.

- *Germany*: a very large article, addressing a large number of different aspects of the country.

- *Hungarian language*: a medium-sized, quite technical article.

- *Pierre de Fermat*: a short biographical article.

- *Star Wars*: a large article, with an important number of links to other articles about the Star Wars universe.

- *Theory of relativity*: a short introductory article, with links to more specialized articles.

- *1989*: a very large article, containing all the important events of year 1989.

It is to be noted that, in order to avoid any bias, we did not run the methods on these articles before the evaluation procedure was launched.

People were asked to assign a mark between 0 and 10 (10 being the best) to the list of the first 20 results returned by each method on these articles, according to their relevance as "related articles" lists. Each evaluator was free to interpret the meaning of the phrase "related articles". The lists were unlabeled, randomly shuffled, and in a potentially different order for each article. The evaluators were allowed to skip articles they did not feel confident enough to vote on.

There has been a total of 67 participants, most of them colleagues and friends of the authors. Not all of them evaluated every article: per-article participation ranges from 41 (*Clique (graph theory)*) to 62 (*Germany*). This allows for reasonably good confidence intervals for the marks, as shown below.

4.4 Performance of the Methods

Table 1 shows the output of **Green** on every evaluation article. Due to lack of space, we only present a portion of the outputs of the other methods in Table 2. The full output and detailed evaluation results can be found in [14].

The average marks given by the evaluators are presented in a radar chart on Figure 1. Each axis stands for the mark given for an article: from worst (0/10) at the center to best (10/10) at the periphery, while each line represents a method

Table 1: Output of Green on the articles used for evaluation.

Clique (graph theory)	*Germany*	*Hungarian language*	*Pierre de Fermat*	*Star Wars*	*Theory of relativity*	*1989*
1. Clique (graph theory) 2. Graph (mathematics) 3. Graph theory 4. Category:Graph theory 5. NP-complete 6. Complement graph 7. Clique problem 8. Complete graph 9. Independent set 10. Maximum common subgraph isomorphism problem 11. Planar graph 12. Glossary of graph theory 13. Mathematics 14. Connectivity (graph theory) 15. Computer science 16. David S. Johnson 17. Independent set problem 18. Computational complexity theory 19. Set 20. Michael Garey	1. Germany 2. Berlin 3. German language 4. Christian Democratic Union (Germany) 5. Austria 6. Hamburg 7. German reunification 8. Social Democratic Party of Germany 9. German Empire 10. German Democratic Republic 11. Bavaria 12. Stuttgart 13. States of Germany 14. Munich 15. European Union 16. National Socialist German Workers Party 17. World War II 18. Jean Edward Smith 19. Soviet Union 20. Rhine	1. Hungarian language 2. Slovakia 3. Romania 4. Slovenia 5. Hungarian alphabet 6. Hungary 7. Croatia 8. Category:Hungarian language 9. Turkic languages 10. Finno-Ugric languages 11. Austria 12. Serbia 13. Uralic languages 14. Ukraine 15. Hungarian grammar (verbs) 16. German language 17. Hungarian grammar 18. Khanty language 19. Hungarian phonology 20. Finnish language	1. Pierre de Fermat 2. Toulouse 3. Fermat's Last Theorem 4. Diophantine equation 5. Fermat's little theorem 6. Fermat number 7. Grandes écoles 8. Blaise Pascal 9. France 10. Pseudoprime 11. Lagrange's four-square theorem 12. Number theory 13. Fermat polygonal number theorem 14. Holographic will 15. Diophantus 16. Euler's theorem 17. Pell's equation 18. Fermat's theorem on sums of two squares 19. Fermat's spiral 20. Fermat's factorization method	1. Star Wars 2. Dates in Star Wars 3. Palpatine 4. Jedi 5. Expanded Universe (Star Wars) 6. Star Wars Episode I: The Phantom Menace 7. Star Wars Episode IV: A New Hope 8. Obi-Wan Kenobi 9. Star Wars Episode III: Revenge of the Sith 10. Coruscant 11. Anakin Skywalker 12. Lando Calrissian 13. Luke Skywalker 14. Star Wars: Clone Wars 15. List of Star Wars books 16. George Lucas 17. Star Wars Episode II: Attack of the Clones 18. Splinter of the Mind's Eye 19. List of Star Wars comic books 20. The Force (Star Wars)	1. Theory of relativity 2. Special relativity 3. General relativity 4. Spacetime 5. Lorentz covariance 6. Albert Einstein 7. Principle of relativity 8. Electromagnetism 9. Lorentz transformation 10. Inertial frame of reference 11. Speed of light 12. Galilean transformation 13. Local symmetry 14. Category:Relativity 15. Galilean invariance 16. Gravitation 17. Global symmetry 18. Tensor 19. Maxwell's equations 20. Introduction to general relativity	1. 1989 2. Cold War 3. 1912 4. Tiananmen Square protests of 1989 5. Soviet Union 6. German Democratic Republic 7. George H. W. Bush 8. 1903 9. Communism 10. 1908 11. 1929 12. Ruhollah Khomeini 13. March 1 14. Czechoslovakia 15. June 4 16. The Satanic Verses (novel) 17. 1902 18. November 7 19. October 9 20. March 14
Mark: 7.6/10	Mark: 7.0/10	Mark: 6.2/10	Mark: 7.3/10	Mark: 7.4/10	Mark: 8.1/10	Mark: 5.4/10

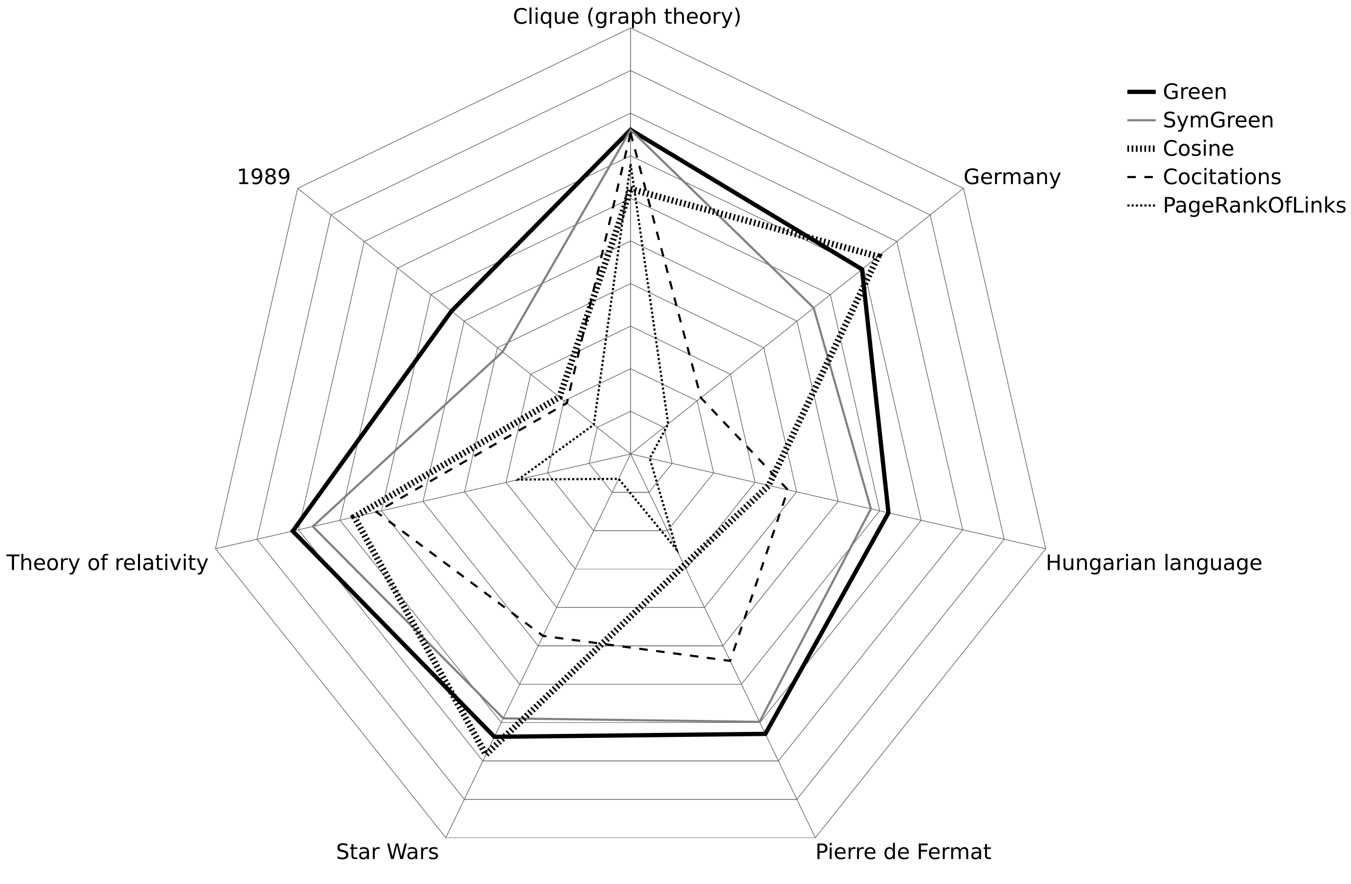

Figure 1: Radar chart of the average marks given to each method on the various base articles.

Table 2: Output of SymGreen, Cosine, Cocitations, and PageRankOfLinks on sample articles.

SymGreen		Cosine		Cocitations		PageRankOfLinks	
Pierre de Fermat	*Germany*	*Pierre de Fermat*	*Germany*	*Pierre de Fermat*	*Germany*	*Pierre de Fermat*	*Germany*
1. Pierre de Fermat 2. Mathematics 3. Probability theory 4. Fermat's Last Theorem 5. Number theory 6. Toulouse 7. Diophantine equation 8. Blaise Pascal 9. Fermat's little theorem 10. Calculus 11. Diophantus 12. Statistics 13. Geometry 14. 17th century 15. Fermat number 16. 1601 17. Mathematician 18. 1665 19. Probability 20. Grandes écoles	1. Germany 2. Berlin 3. France 4. Austria 5. German language 6. Bavaria 7. World War II 8. German Democratic Republic 9. European Union 10. Hamburg 11. Christian Democratic Union (Germany) 12. West Germany 13. Denmark 14. Stuttgart 15. Social Democratic Party of Germany 16. German reunification 17. German Empire 18. States of Germany 19. Munich 20. Switzerland	1. Pierre de Fermat 2. ENSICA 3. Fermat's theorem 4. International School of Toulouse 5. École Nationale Supérieure d'Électronique, d'Électrotechnique, d'Informatique, d'Hydraulique, et de Télécommunications 6. Languedoc 7. Hélène Pince 8. Community of Agglomeration of Greater Toulouse 9. Lilhac 10. Institut d'études politiques de Toulouse 11. Bonhoure Radio Tower 12. École Nationale de la Statistique et de l'Administration Économique 13. Cathédrale Saint-Étienne de Toulouse 14. List of Pink Cities 15. Number theory 16. Battle of Toulouse (1814) 17. Wieferich prime 18. Jean-Baptiste Dortignacq 19. Saint-Jean, Haute-Garonne 20. European Physiology Modules	1. Germany 2. History of Germany since 1945 3. History of Germany 4. Timeline of German history 5. States of Germany 6. Politics of Germany 7. List of Germany-related topics 8. Hildesheimer Rabbinical Seminary 9. Pleasure Victim 10. German Unity Day 11. Gay rights in Germany 12. Wolfgang Becker 13. Kitty-Yo 14. Metrinomics - Metrivox 15. Germans 16. Basic Law for the Federal Republic of Germany 17. Autobahn 18. West Germany 19. German reunification 20. Veolia Verkehr	1. Pierre de Fermat 2. Leonhard Euler 3. Mathematics 4. René Descartes 5. Mathematician 6. Gottfried Leibniz 7. Calculus 8. Isaac Newton 9. Blaise Pascal 10. Carl Friedrich Gauss 11. Number theory 12. Euclid 13. Geometry 14. France 15. Joseph Louis Lagrange 16. Diophantus 17. Fermat's Last Theorem 18. Algebra 19. Archimedes 20. Differential equation	1. Germany 2. United States 3. France 4. United Kingdom 5. World War II 6. Italy 7. Netherlands 8. Japan 9. 2005 10. Category:Living people 11. Canada 12. Spain 13. Poland 14. Austria 15. Russia 16. Australia 17. England 18. 2004 19. Switzerland 20. Europe	1. France 2. 17th century 3. March 4 4. January 12 5. August 17 6. Calculus 7. Lawyer 8. 1660 9. Number theory 10. René Descartes 11. Probability theory 12. Carl Friedrich Gauss 13. 1665 14. Toulouse 15. 1601 16. Blaise Pascal 17. Analytic geometry 18. Geometric progression 19. Parlement 20. Pierre de Fermat	1. United States 2. United Kingdom 3. France 4. 2005 5. Germany 6. World War II 7. Canada 8. English language 9. Japan 10. Italy 11. Europe 12. India 13. Russia 14. Latin 15. London 16. China 17. Soviet Union 18. French language 19. Roman Catholic Church 20. Netherlands
Mark: 7.0/10	Mark: 5.5/10	Mark: 2.9/10	Mark: 7.4/10	Mark: 5.4/10	Mark: 2.1/10	Mark: 2.5/10	Mark: 1.1/10

Table 3: Evaluation results. For each method, the following figures are given: average mark, averaged on all articles; 90% Student's t-distribution confidence interval; article-to-article standard deviation; evaluator-to-evaluator standard deviation; global count of 10/10 marks; average mark for each article.

	Green	SymGreen	Cosine	Cocitations	PageRankOfLinks
Average mark	7.0	6.3	5.2	4.5	2.2
90% confidence interval	±0.3	±0.3	±0.3	±0.3	±0.2
Article std. dev.	0.9	1.3	2.2	1.9	2.0
Evaluator std. dev.	1.7	1.7	1.9	2.0	1.6
Number of 10/10	25	10	12	9	4
Clique (graph theory)	7.6	7.6	6.2	7.5	6.8
Germany	7.0	5.5	7.4	2.1	1.1
Hungarian language	6.2	5.8	3.3	3.8	0.5
Pierre de Fermat	7.3	7.0	2.9	5.4	2.5
Star Wars	7.4	6.9	7.8	4.7	0.6
Theory of relativity	8.1	7.7	6.7	6.1	2.7
1989	5.4	3.8	2.1	1.9	1.1

(cf. the legend). Table 3 gives global statistics about the performance of the methods, namely: Average mark and associated 90% Student's t-distribution confidence interval, averaged on all articles for a given method; Standard deviations between the marks, decomposed as an article-to-article standard deviation and an evaluator-to-evaluator standard deviation; Total number of 10/10 marks each method received (this is an indicator of how often the method can produce "perfect" list of results); Average mark given by the evaluators on each article for each method.

Absolute marks should be taken with caution: it is probable that a human-designed list of related pages would not score very close to 10/10, but maybe closer to 8/10, due to the variations in the ways evaluators attribute marks. The first indication for this is the evaluator-to-evaluator standard deviation on a given article, which is always between 1.5 and 2. Another hint is the performance of Green on *Theory of relativity*: the list was attributed a top 10/10 mark by a significant number of evaluators, including several ones who are experts in this field. This seems to indicate that the output of Green on this particular article is as good as can be expected, yet its average mark is only 8.1/10, due to differences in what evaluators expect.

The first thing to be noted is that Green presents the best overall performance. The difference between global scores of Green and of the best classical approach, Cosine, is 1.8, which is statistically significant since the 90% confidence interval for both values is ±0.3. Moreover, Green comes out first for all but two articles, where it is second with a hardly significant gap (0.4 in both cases). Another strength of Green is its low article-to-article standard deviation (0.9, compared to 2.2 for Cosine and 1.9 for Cocitations), which means, altogether with its good overall performance, that it is very robust. Indeed, a look at Figure 1 shows that it never performs very badly. Another aspect of this robustness is the fact that there are very few completely irrelevant words in its output, as can be seen on Table 1; the high number of 10/10 given to Green is perhaps a measure of this fact. Finally, some of the related articles proposed by Green are both highly semantically relevant and completely absent from the output of other methods: this is the case of *Finnish language* for *Hungarian language* (linguists now consider that both languages are related, which was not obvious from geography and the languages themselves), and of *Tiananmen Square protests of 1989* or *The Satanic Verses (novel)* for *1989*.

Note that *1989* was a kind of worst-case test, due to the number, diversity and lack of selectivity of the links from and to this article, as a glance to the corresponding Wikipedia article shows. Only Green (and, in a lesser way, SymGreen) manage to produce interesting results.

SymGreen presents a similar profile as Green, although it performs a little worse in our evaluation. Comments similar to Green can be made about its robustness or the extraction of relevant semantic information (on *1989*, SymGreen is the only method to output *Berlin Wall*). Actually, on other articles we experimented with in an informal way, SymGreen seems better than Green, and we think that interpolating between the two methods, as described in Section 6, would provide optimal robustness.

Cosine performs best of the "classical" methods, but is clearly not as good as the Green-based ones. It has an interesting behavior in the evaluation, since both very good and very bad performance occur: compare for instance *Germany* and *Pierre de Fermat* in Table 2. The method seems thus to be unstable, which is visible in its high article-to-article standard deviation. Moreover, even in the case when it performs well, as for *Germany*, completely irrelevant or anecdotal entries are proposed as related articles, like *Pleasure Victim* or *Hildesheimer Rabbinical Seminary.* Testing the methods informally on more articles confirmed this serious instability of Cosine: these accidents are not specific to the articles selected for evaluation.

Cocitations does not give very good results, but it is still interesting: more than *related* articles, it outputs lists of *similar* articles, giving for instance names of mathematicians of the same period for *Pierre de Fermat*, languages for *Hungarian language* or years for *1989*.

PageRankOfLinks is the worst of the methods tested (although LocalPageRank, not formally tested here, shows even worse performance). It basically outputs variations on the global PageRank values whatever the base article. The only case when it gives good results is a short and technical article with very few links, *Clique (graph theory)*, on which all methods perform well.

In conclusion, Green (and, in a lesser way, SymGreen) shows three main advantages:

1. The average performance is high, significantly above that of the other methods.

2. It is *robust*, never showing a bad performance on an article.

3. It is able to unveil semantic relationships not found by the other methods.

5. RELATED WORK

To our knowledge, this is the first use of discrete Green measures in the field of information retrieval on graphs or hyperlinked structures.

A lot of work has been performed on the topic of finding related pages on the World Wide Web; in his original well-known paper about *hubs* and *authorities* [11], Kleinberg suggests to use authorities in a focused subgraph in order to compute *similar-page queries*; apart from the use of authorities instead of PageRank, this is very similar to LocalPageRank, which tends to perform very poorly on the Wikipedia graph. In [3], Dean et al. present two different approaches for finding related pages on the Web: the *Companion* algorithm, which uses authorities scores in a local subgraph, and a cocitation-based algorithm. A more original approach is presented by Flake et al. for the identification of *Web communities*: the Web is seen as a traffic network, and a community of Web pages is defined by a maximum flow/minimum cut on this network; this is an interesting direction, but it is more appropriate when there are more than one *seed* Web page to start with (cf. [16]).

Comparatively, the literature on the extraction of *related articles* from Wikipedia is more limited. Adafre et al. [1] use a cocitation approach to identify missing links in Wikipedia. We saw in Section 4.4 that Cocitations fared much worse than Green in our experiment. *Synarcher* [12] is a program for synonym extraction in Wikipedia, relying on authority scores in a local subgraph (comparable to LocalPageRank) together with the information provided by Wikipedia's category structures. Grangier et al. present in [7] a technique which modifies a classical text mining similarity measure of articles (based on the full textual content) by taking the hyperlinks into account using machine learning; no application to the problem of finding related pages is given.

6. PERSPECTIVES

The Green method is rather universal: it takes only a directed graph as its input, and is parameter-free. While this is a strong point of the method, and while it already performs well without adjustment, one might want to use special features of a particular situation at hand to improve performance.

There is much room for variation on the Green method. In the example of Wikipedia, the rich textual content, as well as an intelligent use of the Wikipedia article syntax (which conveys semantic information) and specific filtering schemes (e.g. filtering out *stub* articles) could probably be used to output almost flawless lists of "related articles". Our goal here was only to illustrate the (already satisfying) performance of a "raw" method which could be applied to other situations without a change.

6.1 Application to the Web Graph

On the practical side, the next step is of course to apply the Green method to the graph of the World Wide Web; we did not have the opportunity to do this for practical reasons. It is to be noted, though, that the computing power needed to compute the Green measure centered at a given node is the same as that of computing the PageRank of the graph, something feasible with large clusters of PCs. Because the Web graph has a number of dissimilarities with the Wikipedia graph (in particular, it is much less strongly connected), some differences in the behavior of the Green method are to be expected.

6.2 Theoretical Extensions

On the theoretical side, there are lots of possible directions for improvement of the Green method. First, as mentioned in the discussion of **SymGreen** in section 3.1.2, one possible flaw of the simple Green method is its inability to follow the links backward in order to find related pages. This is corrected in **SymGreen** by taking a linear combination of the forward and backward random walks. While in the case of the seven Wikipedia articles evaluated, the performance of **SymGreen** was slightly worse than the performance of **Green**, we think that in more general situations it is safer to use a version going both forward and backward. If a Markov chain is given by its transition probabilities (p_{ij}), there is a natural interpolation between going forward and backward, depending on a parameter $0 \leqslant \alpha \leqslant 1$. Namely, setting

$$\tilde{p}_{ij} = \alpha\, p_{ij} + (1-\alpha)\, p_{ji}\, \frac{\nu_j}{\nu_i}$$

(where as usual ν_j is the equilibrium measure) defines a Markov chain which goes forward a proportion α of the time and backward a proportion $(1-\alpha)$ of the time. **SymGreen** corresponds to $\alpha = 1/2$, while $\alpha = 1$ for **Green**. Based on informal experiments on more Wikipedia articles, we think that intermediate values of α such as $2/3$ or $3/4$, which favor going forward without forbidding to go backward, could yield better results in general, especially when facing a graph with unknown properties.

Second, we used a weight $\log 1/\nu_j$ to favor "specific" rather than "general" nodes in the output. The difference made by this weight was not overwhelming, however it might be interesting to study and devise better weights.

Third, there is the possibility to merge the cosine and Green methods as follows. The cosine method uses a representation of each node as a vector in $\mathbb{R}^n$ for some n, then evaluates similarity of nodes by computing the angle between corresponding vectors. An interesting possibility would be to use the Green measure centered at node i (which is a row vector) as the vector representation for node i, and then computing cosines on these Green vectors. At present time this is computationally too expensive (since it requires computing all Green measures then comparing them), but it could become practical with more computer resources.

Finally, since PageRank is a simple random walk with random jumps, and the Green method is a random walk with source and sinks, any variation of the PageRank method can be adapted to the Green method. We already discussed (Section 3.2) why adding random jumps would go opposite to the goal of finding related pages. Other algorithms similar to PageRank could have their Green counterparts: for example, adapting the HITS algorithm [11] to define Green measures with hubs and authorities is straightforward.

7. CONCLUSION

We showed how to use Green measures for the extraction of related nodes in a graph. This is a generic, parameter-free algorithm, which can be applied *as is* to any directed graph. We have described and implemented in a uniform way other classical approaches for finding related nodes. Finally, we have carried out a user study on the example of the graph of Wikipedia, which has shown that the Green method significantly outperforms the other methods, is very robust, and is the only one able to find some important semantical relations.

As discussed in Section 6, there is space for extensions and improvements, either on the theoretical or the application side. For example it is easy to design variations on the Green method using standard variations on PageRank. Also, as the methods presented here rely purely on the graph structure, it is likely that, in the specific case of Wikipedia, we can improve performance by taking into account the textual content of the articles, the categories, some templates... Another obvious application is to try the Green method on the Web graph; more generally, it could be directly applied to any other context.

8. ACKNOWLEDGEMENTS

We would like to thank all 67 participants who took the time to evaluate the different methods, and Serge Abiteboul for his feedback on the topic.

9. REFERENCES

[1] S. F. Adafre and M. de Rijke. Discovering missing links in Wikipedia. In *Workshop on Link Discovery: Issues, Approaches and Applications*, Chicago, USA, Aug. 2005.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[3] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–1479, 1999.

[4] D. G. Duffy. *Green's functions with applications.* Studies in Advanced Mathematics. Chapman & Hall/CRC, Boca Raton, FL, 2001.

[5] Google. `http://www.google.com/`.

[6] Google Scholar. `http://scholar.google.com/`.

[7] D. Grangier and S. Bengio. Inferring document similarity from hyperlinks. In *Conference on Information and Knowledge Management*, Bremen, Germany, Nov. 2005.

[8] O. Häggström. *Finite Markov chains and algorithmic applications*, volume 52 of *London Mathematical Society Student Texts.* Cambridge University Press, Cambridge, 2002.

[9] T. Holloway, M. Božičević, and K. Börner. Analyzing and visualizing the semantic coverage of Wikipedia and its authors. `http://arxiv.org/abs/cs.IR/0512085`.

[10] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov chains.* D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1966.

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

[12] A. A. Krizhanovsky. Synonym search in Wikipedia: Synarcher. `http://arxiv.org/abs/cs.IR/0606097`.

[13] J. R. Norris. *Markov chains*, volume 2 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 1997.

[14] Y. Ollivier and P. Senellart. Finding related pages using Green measures: An illustration with Wikipedia, companion website. `http://pierre.senellart.com/wikipedia/`.

[15] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Book Company, 1984.

[16] P. Senellart. Identifying websites with flow simulation. In *International Conference on Web Engineering*, Sydney, Australia, July 2005.

[17] J. Voß. Measuring Wikipedia. In *International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, July 2005.

[18] Wikipedia. The free encyclopedia. `http://en.wikipedia.org/`.

[19] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1), July 2006. Id. 016115.