

Communities in Large Networks: Identification and Ranking

Martin Olsen

Department of Computer Science
University of Aarhus*
mo@daimi.au.dk

Abstract. We study the problem of identifying and ranking the members of a community in a very large network with link analysis only, given a set of representatives of the community. We define the concept of a *community* justified by a formal analysis of a simple model of the evolution of a directed graph. We show that the problem of deciding whether a non trivial community exists is NP complete. Nevertheless, experiments show that a very simple greedy approach can identify members of a community in the Danish part of the web graph with time complexity only dependent on the size of the found community and its immediate surroundings. The members are ranked with a “local” variant of the PageRank algorithm. Results are reported from successful experiments on identifying and ranking Danish Computer Science sites and Danish Chess pages using only a few representatives.

1 Introduction

A community in a network is a set of somewhat isolated vertices linking heavily to each other - for example a set of pages in the web graph related to a particular topic. People controlling a group of vertices (and their outgoing links) in a community are always looking for answers to the questions “How strong are the positions in the community for the members in my group?” and “How can these positions be improved?”. The main objective for the work behind this paper is to establish a model of the community producing satisfactory answers to the first question. The model should also be small enough to enable a formal analysis leading to answers to the second question.

The purpose of the techniques in this paper is not to partition the network in to several communities. The purpose is to isolate and rank the members of a *single* community given by a set of representatives. Before the discussion of related work we would like to introduce the notation used in this paper.

In this paper $G = (V, E)$ denotes a directed graph where multiple occurrences of $(u, v) \in E$ are allowed. We will call $(u, v) \in E$ a *link* on u and say that u links to v etc. A link could for example represent a link from site u to site v in the

* The research is partly sponsored by the Danish company Cofman (www.cofman.com).

web graph or a reference in a paper written by u to a paper written by v . We define the *relative attention* that u shows v as $w_{uv} = \frac{m(u,v)}{\text{outdeg}(u)}$ where $m(u,v)$ is the multiplicity of link (u,v) in E . If $\text{outdeg}(u) = 0$ then $w_{uv} = 0$. For $C \subseteq V$ we let $w_{uC} = \sum_{c \in C} w_{uc}$, i.e. the attention that u shows the set of vertices C . In this paper we will reserve the term *edge* for an undirected graph.

1.1 Related Work

The problem of finding community structures in networks has been subject to a great deal of research - see e.g. [11].

Bagrow *et al.* [3] present a “local” method for detecting the community given by a single representative. A breadth first search from the representative stops when the number of edges connecting the visited vertices with un-visited vertices drops in a special way and reports the visited vertices as a community. Bagrow *et al.* repeat this procedure for each vertex and analyzes the overlap of the communities in order to eliminate problems with what the authors call “spill-over” of the breadth first search.

Formal definitions of communities are provided by Flake and different co-authors in [5] and [6]. According to [5] a community in an *undirected* graph with edges of unit capacity is a set of vertices C such that for all $v \in C$, v has at least as many edges connecting to vertices in C as it does to vertices in $\bar{C} = V - C$. Using the notion of relative attention extended to undirected graphs, this is $\forall v \in C : w_{vC} \geq \frac{1}{2}$. Flake *et al.* show in [5] how to identify a community containing a set of representatives as an s - t minimum cut in a graph with a virtual source s and virtual sink t . They show how the method can process only the neighborhood of the representatives yielding a local method with time complexity dependent on the size of the neighborhood. It is not possible for a vertex within a distance of more than two from the representatives to join the community for this “local” variant of their method.

The web graph is treated as a weighted *undirected* graph in [6] with an edge between page i and page j if and only if there is a link from page i to j or vice versa. Edge $\{i,j\}$ has weight $w_{ij} + w_{ji}$ following our definitions of attention. The graph is expanded with a virtual vertex t connected to all vertices with edges with the same weight α and the *community* of page s is defined by means of an s - t minimum cut. The members of such a community can be identified with a maximum flow algorithm.

The definitions in [5] and [6] are not based on a model of the evolution of a graph. It should also be noted that it seems impossible for a universally popular member to be a member of a small community by the definitions in [5] and [6]. A relatively high in-degree of a member will prevent it from being on the community side of a minimum cut. In fact any member v of a relatively small community in a relatively large network is risking being forced to leave the community if v attracts some attention from non community members if the community definition is based on minimum cuts and the graph is undirected.

Recently Andersen *et al.* [1] and Andersen and Lang [2] presented some very interesting approaches to identifying communities containing specific vertices. In

both papers random walks are used to identify the communities. The graphs are assumed to be *unweighted* and *undirected* where this paper deals with *directed* graphs.

The search engine Google uses the PageRank algorithm [4, 12] to calculate a universal measure of the popularity of the web pages. For a given search query the universal measure is combined with a measure of relevance with respect to the query in order to rank the web pages. Several variants of the PageRank algorithm have been proposed to make it personalized or topic/query specific - see for example [8, 9, 13].

1.2 Our Results

We present a community definition justified by a formal analysis of a very simple model of the evolution of a directed graph. We show that the problem of deciding whether a community $C \neq V$ exists such that $R \subseteq C$ for a given set of representatives R is NP complete. Nevertheless, we show that a fast and simple parameter free greedy approach performs well when detecting communities in the Danish part of the web graph. The time complexity of the approach is only dependent on the size of the found community and its immediate surroundings. Our method is “local” as the method in [3] but it does not use breadth first searches. We also show how to use a computationally inexpensive local variant of PageRank to rank the members of the communities and compare the ranking with the PageRank for the total graph.

These are two possible applications of the algorithms presented in this paper:

- Consider the following scenario: A user interested in Computer Science visits some sites on this subject. A piece of software running in the background finds that the Computer Science sites are similar by analyzing the content of the sites. It uses the Computer Science sites as the set R and reports a community C containing R with the sites ranked by our ranking algorithm. A real world example in Sect. 4.2 documents that this list could be very useful to the user!
- The ranking of the members of a community is the stationary probability distribution of a Markov Chain with the community as the state space. This Markov Chain can form the basis for an analysis leading to answers to questions like “Which link modifications would be optimal wrt. ranking for our group of nodes?”.

In Sect. 2 the community definition and the greedy approach for identifying community members are presented. The ranking algorithm is introduced in Sect. 3 and the experiments are reported in Sect. 4.

2 Locating Communities

2.1 Community Definition

The intuition behind our community definition is that every community member shows more attention to the community than any non member:

Definition 1. A community is a set $C \subseteq V$ such that

$$\forall u \in C, \forall v \in \bar{C} : w_{uC} \geq w_{vC} .$$

Consider the following process: Assume the existence of a set $C \subseteq V$ and numbers p_1 and p_2 with $0 \leq p_1 < p_2 \leq 1$ such that the following holds: Every time a vertex $u \in C$ links to another vertex it will link to a member in C with probability p_2 . Every time a vertex $v \in \bar{C}$ establishes a link it will link to a member in C with probability p_1 . Each member of V establishes exactly q links independently of all other links established.

The number p_2 can be smaller than $\frac{1}{2}$ which means that the members of C does not necessarily predominantly link to other members of C as supposed in [5].

Definition 1 is justified by the following theorem:

Theorem 1. Consider the process defined above and let $n = |V|$. If $\alpha = \left(1 - \frac{p_1}{p_2}\right) / \ln \frac{p_2}{p_1}$ then:

$$P(\forall u \in C, \forall v \in \bar{C} : w_{uC} \geq w_{vC}) \geq 1 - n \left(\frac{e^{\alpha-1}}{\alpha^\alpha} \right)^{p_2 q} . \quad (1)$$

Proof. Let X_{xC} denote the number of links established by x linking to members in C . Let $\mu_2 = p_2 \cdot q$ denote the expected value for X_{uC} if $u \in C$. The expected value for X_{vC} for $v \in \bar{C}$ is $\mu_1 = p_1 \cdot q$.

We will establish an upper bound for the event in (1) *not happening*:

$$\begin{aligned} P(\exists u \in C, \exists v \in \bar{C} : X_{uC} < X_{vC}) &\leq \\ P(\exists u \in C : X_{uC} < \tau \vee \exists v \in \bar{C} : X_{vC} > \tau) &\leq \\ |C| \cdot P(X_{uC} < \tau) + |\bar{C}| \cdot P(X_{vC} > \tau) &. \end{aligned} \quad (2)$$

where u and v are generic elements in C and \bar{C} respectively. This upper bound holds for any value of τ . The strategy of the proof is to find a τ such that the factors $P(X_{uC} < \tau)$ and $P(X_{vC} > \tau)$ have a low common upper bound.

We will use two Chernoff bounds and produce upper bounds on the factors in (2) assuming $\tau = \alpha \mu_2 = \frac{p_2}{p_1} \alpha \mu_1$ for $\alpha \in (\frac{p_1}{p_2}, 1)$:

$$P(X_{uC} < \alpha \mu_2) \leq e^{-\mu_2} \left(\frac{e^\alpha}{\alpha^\alpha} \right)^{\mu_2} . \quad (3)$$

$$P\left(X_{vC} > \frac{p_2}{p_1} \alpha \mu_1\right) \leq e^{-\mu_1} \left(\frac{e}{\frac{p_2}{p_1} \alpha} \right)^{\frac{p_2}{p_1} \alpha \mu_1} = e^{-\mu_1} \left(\frac{p_1}{p_2} \right)^{\alpha \mu_2} \left(\frac{e^\alpha}{\alpha^\alpha} \right)^{\mu_2} . \quad (4)$$

Now we will find a necessary and sufficient condition for these upper bounds to be identical:

$$e^{-\mu_2} = e^{-\mu_1} \left(\frac{p_1}{p_2} \right)^{\alpha \mu_2} \Leftrightarrow$$

$$\begin{aligned}
-\mu_2 &= -\mu_1 + \alpha\mu_2 \ln \frac{p_1}{p_2} \Leftrightarrow \\
\alpha &= \left(1 - \frac{p_1}{p_2}\right) / \ln \frac{p_2}{p_1} .
\end{aligned}$$

The upper bounds in (3) and (4) are identical for this value of α which is easily shown to satisfy $\alpha \in (\frac{p_1}{p_2}, 1)$. We will put the common value $(\frac{e^{\alpha-1}}{\alpha^\alpha})^{\mu_2}$ in (2):

$$P(\exists u \in C, \exists v \in \bar{C} : X_{uC} < X_{vC}) \leq n \left(\frac{e^{\alpha-1}}{\alpha^\alpha}\right)^{p_2 q} .$$

□

Theorem 1 shows that real communities with $p_2 > p_1$ probably will obey Definition 1 in a large network where the number of links from each vertex is logarithmically lower bounded as pointed out by the following corollary:

Corollary 1. *For fixed p_1 and p_2 with $p_1 < p_2$ there exists a constant $k > 0$ such that*

$$P(\forall u \in C, \forall v \in \bar{C} : w_{uC} \geq w_{vC}) \rightarrow 1 \quad \text{for } n \rightarrow \infty .$$

for $q = k \cdot \log n$.

Before addressing computability issues a couple of remarks on our community definition are in place. First of all there might be several communities containing a given set of representatives so picking the representatives might require several attempts. The experiments in Sect. 4.1 deal with the problem of choosing representatives. Secondly the union $C = C_1 \cup C_2$ of two communities C_1 and C_2 is not necessarily a community. For example there might be a vertex $v \in \bar{C}$ with $w_{vC} = 1$ and a vertex $u \in C$ with $w_{uC} < 1$ in which case C would not be a community since $w_{uC} < w_{vC}$. Communities in the “real world” seem to share these properties with our formal communities.

2.2 Intractability

We will now formally define the problem of deciding whether a non trivial community exists for a given set R :

Definition 2. *The COMMUNITY problem:*

- *Instance:* A directed graph $G = (V, E)$ and a set of vertices $R \subset V$.
- *Question:* Does a community $C \neq V$ according to Definition 1 exist such that $R \subseteq C$?

If we had an effective algorithm locating a non trivial community if at least one such community existed then we also could solve COMMUNITY effectively but even solving COMMUNITY effectively seems hard according to the following theorem:

Theorem 2. *COMMUNITY is NP complete.*

Proof. We can check in polynomial time whether C is a community containing R by calculating w_{xC} for all $x \in V$ thus COMMUNITY is in NP.

We will transform an instance of the NP complete problem PARTITION [7, page 223] into an equivalent instance of COMMUNITY in polynomial time. This means that we can solve the NP complete problem PARTITION in polynomial time if we can solve COMMUNITY in polynomial time thus COMMUNITY is NP complete since it is a member of NP. The rest of the proof contains the details of the transformation.

An instance of PARTITION is a finite set $A = \{a_1, a_2, \dots, a_n\}$ and a size $s(a_i) \in \mathbb{Z}^+$ for each $a_i \in A$. The question is whether a subset $A' \subset A$ exists such that $\sum_{a \in A'} s(a) = \frac{S}{2}$ where S is the sum of the sizes of all elements in A ? We will transform this instance into the instance of COMMUNITY given by a directed graph $G(V, E)$ with $n + 2$ vertices and $R = \{r\}$ where r is one of the vertices in G . The graph G is constructed in the following way:

We will start with two vertices r and y . For each $a_i \in A$ we will make a vertex with two links (a_i, r) and (a_i, y) with multiplicity 1 and two links (r, a_i) and (y, a_i) with multiplicity $s(a_i)$. The resulting network is shown on Fig. 1.

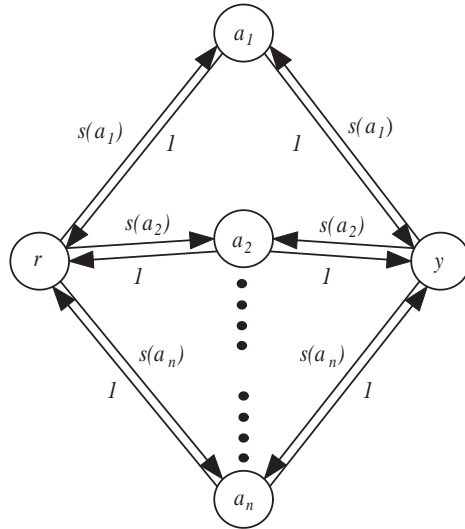


Fig. 1. A non trivial community C with $r \in C$ exists if and only if it is possible to divide the set A in two parts with the same size. Each link is labeled with its multiplicity.

Now we will prove that G contains a non trivial community C containing R if and only if A' exists.

- If A' exists then $C = \{r\} \cup A'$ is a non trivial community containing r since $w_{xC} = \frac{1}{2}$ for all $x \in V$.
- Now assume that C is a non trivial community containing r . If C contains y then C also contains all the a 's since $w_{aC} = 1$ if $\{r, y\} \subseteq C$. Since C is a non trivial community we have $y \notin C$. Now set $A' = C \cap A$.
 - If $\sum_{a \in A'} s(a) < \frac{S}{2}$ then $w_{rC} < \frac{1}{2}$ but there is at least one $a \notin C$ with $w_{aC} = \frac{1}{2}$ contradicting that C is a community.
 - If $\sum_{a \in A'} s(a) > \frac{S}{2}$ then $w_{yC} > \frac{1}{2}$ but there is at least one $a \in C$ with $w_{aC} = \frac{1}{2}$ - yet another contradiction.

We can conclude that $\sum_{a \in A'} s(a) = \frac{S}{2}$. □

The network in Fig. 1 might be illustrative when comparing the definitions of a community in this paper and in [6]. If $A' \subset A$ exists such that $\sum_{a \in A'} s(a) = \sum_{a \in A-A'} s(a)$ then $C = \{r\} \cup A'$ will not be a community by the definition in [6] for any value of α .

2.3 A Greedy Approach

Despite the computational intractability experiments show that it is possible to find communities in the Danish part of the web graph with a simple greedy approach (see Sect. 4).

The approach starts with $C = R$. It then moves one element from \bar{C} to C at a time choosing the element $v \in \bar{C}$ with the highest value of w_{vC} . After moving v to C it updates w_{xC} for all x linking to v and checks whether the current C satisfies Definition 1. The approach can be effectively implemented using two priority queues containing the elements in C and the elements in \bar{C} linking to C respectively using w_{xC} as the key for x . The C -queue is a min-queue and the \bar{C} -queue is a max-queue. It is possible to find the next element to move and to decide if C is a community by inspecting the first elements in the queues as can be seen from the pseudo code of the approach shown in Fig. 2.

The time complexity of the approach is $O((n_C + m_C) \log n_C)$ where n_C is the number of elements in the union of the found community C and the set of vertices linking to C and m_C is the number of links between elements in C plus the number of links to C from \bar{C} - multiple occurrences of $(u, v) \in E$ only counts as one link. The argument for the complexity is that less than n_C elements have to move between the two queues and that m_C update-priority operations are performed on the two queues containing no more than n_C elements. We are assuming that finding one vertex x linking to v can be done in constant time.

Some of the representatives might have no links, so we do not consider the attention shown by the representatives to C when we check whether C satisfies our definition of a community for the experiments in this paper. To be more specific we check whether

$$\forall u \in C - R, \forall v \in \bar{C} : w_{uC} \geq w_{vC} .$$

```

Greedy( $G, R$ )
 $C$ -queue :=  $\emptyset$ 
 $\bar{C}$ -queue :=  $\emptyset$ 
forall  $r \in R$  do
  forall  $x \in V - R$  linking to  $r$  do
    if  $x \in \bar{C}$ -queue then
      increase the priority of  $x$  with  $w_{xr}$ 
    else
      insert  $x$  in the  $\bar{C}$ -queue with priority  $w_{xr}$ 
  while  $|C$ -queue  $<$  minimum size or  $\min(C$ -queue)  $<$   $\max(\bar{C}$ -queue) do
    move the element  $v$  with maximum priority from the  $\bar{C}$ -queue to the  $C$ -queue
  forall  $x \in V - R$  linking to  $v$  do
    if  $x \in C$ -queue or  $x \in \bar{C}$ -queue then
      increase the priority of  $x$  with  $w_{xv}$ 
    else
      insert  $x$  in the  $\bar{C}$ -queue with priority  $w_{xv}$ 
  Report  $R \cup C$ -queue as a community

```

Fig. 2. Pseudo code for the greedy approach. Details for handling an empty C -queue or an empty \bar{C} -queue in the while-loop have been omitted for clarity.

3 Ranking the Members

The PageRank algorithm used by Google can be viewed as a vote among *all* pages yielding a global measure of popularity. We will turn this into a vote among the *relevant* pages that are the pages in C . In this way we will obtain a small mathematical model which forms a basis for analyzing the consequences of changes in the link structure. The experiments carried out produces what we believe to be very valuable rankings which support the validity of the mathematical models behind the rankings.

A visitor to a community member $i \in C$ is assumed to have the following behavior:

- With probability given by some number d he decides to follow a link¹ from i . In this case there are two alternatives:
 - He decides to visit another member j of C . The probability that j gets a visit in this way is $d \cdot w_{ij}$.
 - He follows a link to a non member v . Assuming a low upper bound on w_{vC} it is not likely that the visitor will use a link to go back to C . Thus we treat this case as a jump to another member of C chosen uniformly at random.
- With probability $1 - d$ he decides to jump to another place without following a link which is treated as a jump to a member in C chosen uniformly at random.

¹ As suggested in [4] we use $d = 0.85$

A visitor to $i \in C$ will visit $j \in C$ with probability

$$p_{ij} = \frac{1-d}{|C|} + \frac{d(1-w_{iC})}{|C|} + d \cdot w_{ij} = \frac{1-d \cdot w_{iC}}{|C|} + d \cdot w_{ij} .$$

Like PageRank the ranking of the members is simply the unique stationary probability distribution of the Markov chain given by the transition matrix $P = \{p_{ij}\}_{i,j \in C}$. An iterative calculation of $\pi \cdot P^i$ will converge to the ranking in a few iterations where π is an arbitrary initial probability distribution. For details on convergence rates etc. we refer to the work of Langville and Meyer [10].

4 Experimental Work

For an on-line version of the results of the experiments please visit the home page of the author: www.daimi.au.dk/~mo/. Besides the results reported in this paper you can also find results from experiments with the *s-t* minimum cut approach from [6].

4.1 Identification of Community Members in Artificial Graphs

Inspired by Newman *et al.* [11] we test the greedy approach on some random computer generated graphs with known community structure. The graphs contain 128 vertices divided into four groups with 32 vertices each with vertices 1 - 32 in the first group, 33 - 64 in the next group etc. We will denote the first of the four groups as *group 1*. For each pair of vertices u and v either two links - (u, v) and (v, u) - or none are added to the graph. The pairs of links are placed independently at random such that the *expected* number of links from a vertex to vertices in the same group is 9 and the expected number of links to vertices outside the group is 7.

For 10 graphs the greedy approach reported the first community found containing at least 32 members with vertex number 1 as the single representative. The average size of the community found was 64.3 and the average number of vertices from group 1 in the community found was 28.9. If we use vertices 1 to 5 as representatives instead the corresponding numbers are 39.3 and 31.3 and if we use vertices 1 to 10 as representatives the numbers are 32.4 and 31.2. These admittedly few experiments suggest that the greedy approach can actually identify members of communities if the number of representatives is sufficient.

4.2 Identification and Ranking of Danish Computer Science Sites

Now we will demonstrate that the greedy approach is able to identify communities in the web graph using only a few representatives. A crawl of the Danish part of the web graph from April 2005 was used as the basis for the web experiments. In the first experiment reported in this paper V consists of the 180468 *sites* in the crawl where a link from site u to v is represented by $(u, v) \in E$.

The objective of the experiment was to identify and rank Danish Computer Science sites. The following four sites were used as representatives:

Table 1. The Top 20 of two communities of Danish Computer Science sites. Representatives are written with bold font. The numbers after a site is the “global” ranking in the dk domain.

	556 members	1460 members
1	www.daimi.au.dk 267	www.au.dk 109
2	www.diku.dk 655	www.sdu.dk 108
3	www.itu.dk 918	www.daimi.au.dk 267
4	www.cs.auc.dk 1022	www.hum.au.dk 221
5	www.brics.dk 1132	www.diku.dk 655
6	www.imm.dtu.dk 1124	www.ifa.au.dk 681
7	www.dina.kvl.dk 1153	www.itu.dk 918
8	www.agrsci.dk 1219	www.ruc.dk 945
9	www.foejo.dk 1504	www.phys.au.dk 1051
10	www.darcof.dk 2113	www.brics.dk 1132
11	www.it-c.dk 2313	www.cs.auc.dk 1022
12	www.dina.dk 2169	www.dina.kvl.dk 1153
13	www.cs.aau.dk 2010	www.imm.dtu.dk 1124
14	rapwap.razor.dk 4585	www.agrsci.dk 1219
15	imv.au.dk 2121	www.kvinfo.dk 1122
16	razor.dk 2990	www.foejo.dk 1504
17	www.imada.sdu.dk 2998	www.bsd-dk.dk 1895
18	www.plbio.kvl.dk 3543	www.humaniora.sdu.dk 1826
19	www.math.ku.dk 2634	www.imv.au.dk 2121
20	mahjong.dk 3813	www.statsbiblioteket.dk 867

- **www.itu.dk**, IT University of Copenhagen
- **www.cs.auc.dk**, Department of Computer Science, University of Aalborg
- **www.imm.dtu.dk**, Department of Informatics and Mathematical Modeling, Technical University of Denmark
- **www.imada.sdu.dk**, Department of Mathematics and Computer Science, University of Southern Denmark

The sites of the Departments of Computer Science for the two biggest universities in Denmark, **www.diku.dk** and **www.daimi.au.dk**, were *not included* in the set of representatives. These sites represent the universities in Copenhagen and Aarhus respectively.

The greedy approach found several communities. The Top 20 ranking of two communities with 556 and 1460 sites respectively are shown in Table 1 which also shows the ranking produced by a PageRank calculation on the dk domain. Members of both communities use more than 15-16% of their links to other members and non members use less than 15-16% on members.

The Top 20 lists contain mainly academic sites and the smaller community seems to be dominated by sites related to Computer Science. The ranking seems to reflect the “sizes” of the corresponding real world entities. It is worth noting that **www.daimi.au.dk** and **www.diku.dk** are ranked 1 and 2 in the smaller community. The site ranked 5 in the smaller community represents BRICS, Basic

Research in Computer Science, which is an international PhD school within the areas of computer and information sciences, hosted by the Universities of Aarhus and Aalborg.

The larger community seems to be a more general academic community with the sites for University of Aarhus and University of Southern Denmark ranked 1 and 2 respectively. The larger community obviously contains the smaller community by the nature of the greedy approach.

The local ranking seems to reflect the global ranking with a few exceptions. The site rapwap.razor.dk is popular among the relevant sites but seems not to be that popular overall. The person behind rapwap.razor.dk has pages in Top 5 on Google searches² for Danish pages on “cygwin” and “php” which justifies rapwap.razor.dk’s place on the Top 20 list of Danish Computer Science sites.

4.3 Identification and Ranking of Danish Chess Pages

We also carried out an experiment at the *page level* in order to rank Danish Chess pages using *one representative only*: www.dsu.dk, the homepage for the Danish Chess Federation. For this experiment V consisted of all pages up to three inter site links away from the representative where the links were considered unoriented. V contains approximately 330.000 pages. The weight w_{uv} is the fraction of inter site links on page u linking to page v .

The greedy approach located a community with 471 members. All members use at least 1.4 % of their inter site links on members and non members use less than 1.4 % on members. This means that only heavily linked non members link to the pages in the community and if they do they only link to the community with a few links. The Top 20 for this experiment – using the ranking from Sect. 3 – is shown in Table 2.

The page ranked 2 in the Top 20 is a page for a chess tournament held in Denmark in 2003 with several Grandmasters competing. The pages ranked 13 and 20 are pages (at that time) for the Danish and Scandinavian Chess championships respectively. Several of the subdivisions of the Danish Chess Federation (4, 7, 9, 19) are represented on the Top 20 and the page ranked 6 provides access to a database of more than 40.000 Chess games³. Most of the rest of the pages on the Top 20 are Chess Club pages. All in all the Top 20 seems useful from a Danish chess players point of view.

For comparison we searched Google⁴ for Danish pages containing the word “skak” – the Danish word for chess. Several of the sites with pages in the Top 20 from Table 2 are also present in the Google search result but the latter seems targeted at a broader chess audience. The Google Top 20 contains for example several pages dealing with on-line chess and chess programs. The Top 20 from Table 2 seems to be targeted at a dedicated Danish chess player being a member of a chess club.

² The searches were carried out on January 23 2007.

³ Appear to have moved to <http://dsu9604.dsu.dk/partier/danbase.htm>.

⁴ The searches were carried out on April 12 2007.

Table 2. The top 20 of a community of 471 Danish chess pages found with the homepage of the Danish Chess Federation as a representative (written with bold font). The Danish word for chess is “skak”.

1.	www.dsu.dk
2.	www.sis-mh-masters.dk
3.	dsus.dk
4.	www.8-hk.dk
5.	www.dsus.dk
6.	www.dsu.dk/partier/danbase.htm
7.	www.vikingskak.dk/4hk
8.	www.sk1968.dk
9.	www.4hk.dk
10.	www.skovlundeskakklub.dk
11.	www.vikingskak.dk
12.	www.alssundskak.dk
13.	www.skak-dm.dk
14.	www.frederikssundskakklub.dk
15.	www.birkeskak.dk
16.	home13.inet.tele.dk/dianalun
17.	www.rpiil.dk/nvf
18.	www.enpassant.dk/chess/index.html
19.	www.4hk.dk/index.htm
20.	www.skak-nm.dk

Acknowledgments The author of this paper would like to thank Torsten Suel and his colleagues at Polytechnic University in New York for a crawl of the Danish part of the web graph and Gerth S. Brodal from University of Aarhus for valuable comments and constructive criticism.

References

1. Reid Andersen, Fan R. K. Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *FOCS*, pages 475–486. IEEE Computer Society, 2006.
2. Reid Andersen and Kevin J. Lang. Communities from seed sets. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 223–232, New York, NY, USA, 2006. ACM Press.
3. Jim Bagrow and Erik Bollt. A local method for detecting communities. *Physical Review E*, 72:046108, 2005.
4. Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.
5. Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 20–23 2000.
6. Gary Flake, Robert Tarjan, and Kostas Tsioutsoulis. Graph clustering and minimum cut trees. *Internet Mathematics*, 1(4):385–408, 2004.
7. M.R. Garey and D.S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

8. Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM Press.
9. Glen Jeh and Jennifer Widom. Scaling personalized web search. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA, 2003. ACM Press.
10. Amy N. Langville and Carl D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2005.
11. M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.
12. Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
13. Mathew Richardson and Pedro Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.