Applications of Page Ranking in P Systems

Michael Muskulus

Mathematical Institute, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
muskulus@math.leidenuniv.nl

Abstract. The page rank of a webpage is a numerical estimate of its authority. In Google's PageRank algorithm the ranking is derived as the invariant probability distribution of a Markov chain random surfer model. The crucial point in this algorithm is the addition of a small probability transition for each pair of states to render the transition matrix irreducible and aperiodic. The same idea can be applied to P systems, and the resulting invariant probability distribution characterizes their dynamical behavior, analogous to recurrent states in deterministic dynamical systems. The modification made to the original P system gives rise to a new class of P systems with the property that their computations need to be robust against random mutations. Another application is the pathway identification problem, where a metabolite graph is constructed from information about biochemical reactions available in public databases. The invariant distribution of this graph, properly interpreted as a Markov chain, should allow to search pathways more efficiently than current algorithms. Such automatic pathway calculations can be used to derive appropriate P system models of metabolic processes.

1 Introduction and Background

Page ranking is the process of assigning a quantitative measure of "authority" to a webpage. Internet search engines usually use a combination of key word related measures and general page ranks to order the results of a user query. These results are displayed in a linear order, and the higher the rank of a webpage, the higher in the resulting list it is displayed. Since a higher rank means a higher visibility, there has developed a large commercial interest in optimizing a webpage's content with the goal of improving its ranking, and nowadays the activity of *search engine optimization* has become a full-time job for many people.

On the one hand, users of a search engine expect results that lead them to their desired search goals efficiently, so in a way a search engine should optimize their ranking methods with regards to user preferences. In particular, it can be argued that a search engine should use ranking strategies which are objective and unbiased. But note that this leads to a dilemma: if a search engine would openly publish its ranking algorithms, on the one hand this would benefit its users, since then they could, in principle at least, target their queries better.

On the other hand, this knowledge would enable owners and designers of webpages to target their desired audience by specific search engine optimization

D. Corne et al. (Eds.): WMC9 2008, LNCS 5391, pp. 311–324, 2009.

strategies — which might not be what users desire. At the moment, search engines therefore keep their algorithms and ranking methods as closely guarded secrets. This is, of course, not the only possible solution, but seems to also stem from (i) considerations about competition between distinct search engines, and (ii) probably the assumption that the benefit for the common user would be negligible, since on the average s/he would not be able to understand the algorithms, whereas commercial companies would.

A particular case is Google, probably the most important general purpose search engine of today. It is believed by professional consultants that its page ranking methods take into account more than 200 distinct factors[1], but Google states that the "heart of their software" is an algorithm called *PageRank* [16], whose name seems to be inspired by the last name of Google founder Lawrence Page [38].

The original ranking algorithm behind Google has been published [8,33] and is also patented (sic!) as a "Method for node ranking in a linked database" (US patent no. 6.285.999), assigned to Stanford University. It can be shown that *PageRank* is natural in the sense that a few axioms, motivated by the theory of social choice, uniquely characterize *PageRank* [2]. Interestingly, the same method has recently been proposed as a new method of citation analysis that is more authoritative, as self-citations have less impact than in traditional citation analysis [28].

In the following we will describe applications of page ranking in the area of membrane systems [34,35]. We will specifically concentrate on the original *PageRank* algorithm, since it is closely related to Markov chain modeling of dynamical P systems as in [31]. The applications that we will discuss are (i) defining the *recurrent behavior* of dynamical P systems, which results in (ii) a new *complexity measure* for dynamical P systems; (iii) proposing a new class of P systems with interesting robustness properties, and (iv) discussing applications in the *identification* of P systems, where biochemical databases are used to infer P system models via pathway extraction.

2 The PageRank Algorithm

The description of the *PageRank* algorithm is usually given in terms of the so-called webgraph. This is the directed graph $D = (V, A)$ where each node $u \in V$ represents a webpage and each arc $(u, v) \in A \subseteq V^2$ represents a link. A link from page $u \in V$ to $v \in V$ can be thought of as providing evidence that v is an "important" page or, more generally, as a *vote* for page v. Intuitively, the more authoritative page u itself is, the higher its vote for page v should count, leading to a recursive definition as follows. Let

$$\begin{aligned} r: \quad & V \to \mathbb{R}_+ \\ & v \mapsto r(v) \end{aligned}$$

[1] An analysis of the most important factors used by Google can be found on `http://www.seomoz.org/article/search-ranking-factors`

be the *ranking function* that assigns a numerical value $r(v)$ to each node v in the webgraph. Then

$$r(v) := \sum_{u \in \{V \mid (u,v) \in A\}} \frac{r(u)}{\mathrm{outdeg}(u)}$$

where the sum runs over all nodes u linking to the page v and $\mathrm{outdeg}(u)$ is the out-degree of node u. In the above interpretation, each page thus transfers its own *PageRank* value equally to all of its link targets. Note that webpages can link multiple times to the same page, but that this counts as only one link, i.e., one arc in the webgraph.

To see that the *PageRank* ranking function is well defined, we need to turn to the theory of Markov chains [7]. Since the webgraph is finite, the function r can be normalized such that $\sum_{v \in V} r(v) = 1$. One can then interpret $r \in \mathbb{R}_+^{|V|}$ as a probability distribution over the set V of webpages. The *transition matrix*

$$P_{uv} = \begin{cases} 1/\mathrm{outdeg}(u) & \text{if } (u,v) \in A, \\ 0 & \text{if } (u,v) \notin A \end{cases}$$

then corresponds to the model for a person surfing between web pages, from now on simply addressed as a *surfer*, as described in [8]. In this so-called *random surfer model* a surfer is considered who randomly follows links, without any preference or bias. The matrix P_{uv} then describes the probability for the surfer, being at page u, to visit page v next. The *PageRank* definition is then equivalent to the following matrix equation:

$$r = P^t r.$$

In the language of Markov chain theory this means that r is required to be a *stationary distribution*. In other words, if a large number of random surfers find themselves, at the same time, at webpages distributed according to the probability distribution r, then after randomly following a link, the individual surfers would end up at different pages, but the number of surfers visiting each webpage would stay approximately the same (exactly the same in the limit of an infinite number of surfers).

Markov chain theory tells us when such a stationary distribution exists and when it is unique. By the ergodic theorem for Markov chains, an *aperiodic* and *irreducible* transition matrix P is sufficient. The transition matrix is aperiodic if the least common multiple of all possible circuits in the webgraph is trivial. This can always be assumed for general digraphs, since only very special digraphs are periodic. Irreducibility is the requirement that each webpage is reachable from each other page, i.e., that the webgraph is strongly connected, and this is usually *not* fulfilled by the transition matrix. In particular, the webgraph usually has pages without outbound links, so-called *dangling* pages or, in the language of Markov chain theory, *sinks* or *black holes*. If one were to apply the *PageRank* idea to a digraph with one or more of these, they would effectively absorb all probability, since eventually a random surfer would always end up in a black hole and stay there forever. To be more precise: One would expect that the resulting invariant distribution would be zero for all non-sinks, and each sink would be assigned the probability of ending up in it, starting from a random page, in accordance with the random surfer model. However, this is not true. There simply would not exist any stationary distribution in such a case. This "singular" behavior led some people to call such nodes black holes, since the usual laws of Markov chain theory cease to work when one of these is encountered.

The solution to this problem is the truly original idea of the founders of Google: In analogy with the random surfer model, it is assumed that a surfer ending on a sink gets bored and turns *randomly* to a new page from the whole webgraph, which is called *teleportation* in [20]. Of course, this is a somewhat unrealistic model for actual internet user behavior, since how does a surfer *find* a *random* webpage (and with uniform probability)? But changing the transition matrix accordingly,

$$\bar{P}_{uv} = \begin{cases} 1/\mathrm{outdeg}(u) & \text{if } (u,v) \in A, \\ 1/|V| & \text{if } \mathrm{outdeg}(u) = 0, \\ 0 & \text{if } \mathrm{outdeg}(u) > 0 \text{ and } (u,v) \notin A \end{cases}$$

leads to a matrix with irreducible *blocks* (which is still not irreducible, except in special cases). Finally, extending this idea and assuming that the surfer has a certain chance $\alpha > 0$ of turning to a random page *every* time s/he follows a link, leads to

$$\bar{\bar{P}}_{uv} = \begin{cases} 1/|V| & \text{if } \mathrm{outdeg}(u) = 0, \\ \alpha/|V| & \text{if } \mathrm{outdeg}(u) > 0 \text{ and } (u,v) \notin A, \\ \alpha/|V| + (1-\alpha)/\mathrm{outdeg}(u) & \text{if } (u,v) \in A \end{cases} \tag{1}$$

which is truly an irreducible and aperiodic matrix [26]. The stationary distribution r is then, also by the ergodic theorem, an *asymptotic distribution.* This means that a random surfer, starting at an arbitrary webpage, has the chance $r(u)$ to be at page $u \in V$, if he has followed a large number of links, using P_{uv} as transition matrix:

$$\lim_{n\to\infty} (P^t)^n x_0 = r, \tag{2}$$

independent of the initial distribution x_0, i.e., his/her starting page. Note that there is a probability $\alpha/|V|$ that the random surfer *stays* at the same page (we can also say that the surfer *accidentally* jumps to the same page that he comes from), i.e., we explicitly allow self-transitions here, since it makes the mathematical analysis simpler.

These results are consequences of the Perron-Frobenius theorem [5], which also shows that r is the (normalized) *dominant eigenvector* of P^t, i.e., the corresponding eigenvalue $\lambda = 1$ is the largest eigenvalue P^t possesses. In practice, the direct computation of the dominant eigenvector for the (sparse) transition matrix of the webgraph is very difficult, due to the graph's enormous size. On the other hand, Eq. 2, starting from the uniform distribution $x_0(u) = 1/|V|$ is used in practice, and is usually called the *power method* [15]. See [20] for further improvements.

3 P Systems and the Random Surfer Model

P system is a general term to describe a broad class of unconventional models of computation that are usually based on multiset rewriting in a hierarchical structure of so-called membranes [35], but also include computational models based on other mechanisms, for example string or grammar rewriting. Originally introduced by Gheorghe Păun in a seminal paper [34], nowadays there exists a large community of researchers working on and with different extensions and variants of P systems.

How are we to interpret the above changes in the context of P systems, i.e., when we are thinking about the random surfer model with possible jumps (Eq. 1) not only as mathematically sufficient and convenient, but rather as a feature of a P system? Obviously, such a mechanism can turn the multisets that describe the object content of a P system into completely different multisets – and we need to control the outcome of such an operation somehow. Before discussing the problems and possible solutions, the following example illustrates the notion of an invariant distribution in a simple class of P systems.

Let us consider the case of a *probabilistic P system* as in [11]. Starting from an initial configuration (multiset) c_0 the evolution of a probabilistic P system generates a rooted tree S of possible states, where each state $i \in S$ is encountered with a probability $p_{c_0,i}$ during the computation (confer [35]). The leaves $\mathcal{L} \subset S$ of this tree are the halting states and each halting state $h \in \mathcal{L}$ is reached with a probability p_h, where $\sum_{h \in \mathcal{L}} p_h = 1$. If we now introduce additional transitions from each halting state back to the initial state c_0, the state space has the structure of an irreducible Markov chain. This Markov chain could be periodic, but it is easy to see that for a such a finite "closed tree" an invariant probability distribution exists as in the case of an irreducible and aperiodic Markov chain. In fact, the unique invariant distribution is given by $\mu_i = 1/|S| \cdot p_{c_0,i}$ for each state $i \in S$. The factor $1/|S|$ has been introduced such that $\sum_{i \in S} \mu_i = 1$.

We see that the concept of invariant distribution generalizes the probability of reaching a halting state in a probabilistic P system. However, the requirement that the evolution has a tree structure makes the class of probabilistic P systems very special. Moreover, the evolution of such a system is not the same as the dynamics described by Eq. 1.

We consider *flat* P systems in the following, i.e., P systems with exactly one membrane. It is well known that each static[2]P system with k membranes is isomorphic to a flat P system, so this is no restriction.

Problem 1. The state space of P systems is usually not known a priori.

Remark 1. Generating the state space of a P system corresponds to the well known *reachability problem*. But if the P system operates in the *asymptotic regime*, i.e., when there are enough objects in the system such that all rules are

[2] A P system is static if the membrane structure does not evolve in the course of time. To be more precise: membrane creation or destruction are not allowed in a static P system.

applicable, it can be considered a vector-addition system on the infinite lattice $\mathbb{Z}^n$, where n is the number of distinct objects. The state space is the affine image of $\mathbb{R}^m$ (m being the number of rules) under the stoichiometric map M, which for a given initial condition $c \in \mathbb{Z}^n$ is the sublattice $c + M\mathbb{R}^m$ of $\mathbb{Z}^n$. In this case, the geometry of the state space is easy to understand and reachability can be efficiently tested [31].

In general P systems, we are often only interested in a finite subset $Q \subset S$ of state space S, and the restriction of the invariant distribution to Q. Due to Eq. 2 the invariant distribution on Q can be approximated by *simulating* the P system a large number of times, *provided* that it is aperiodic and irreducible.

Problem 2. The state space of P systems is usually infinite.

Remark 2. This is a variant of the previous problem. In principle, for a countably infinite state space there can still exist invariant distributions (see [7] for results about when this is known to be the case). However, some problematic issues surface with an infinite state space for the teleportation property. In particular, the probability of teleportating from state $i \in V$ to some state $j \in V$ is equal to zero[3] when $|V| = \infty$. The only solution of this problem is to somehow modify the teleportation property (confer Section 4).

Problem 3. P systems are non-deterministic, and not probabilistic. What sense does an invariant probability distribution make for a non-deterministic system?

Remark 3. The easy solution of this problem is to only consider variants of P systems that are probabilistic instead of being non-deterministic (see Section 5).

However, for the sake of the argument, let us consider a truly non-deterministic system. It can be considered as the equivalence class of all probabilistic systems with non-zero transition probabilities $P_{ij} > 0$ exactly for all states $i, j \in V$, where j is reachable from i in one time step. An invariant distribution has the property that its support is the whole state space, so the notion of invariant distribution for a non-deterministic system is equivalent with the information what the state space is. Note that quantitative information *can* be obtained in non-deterministic systems (see the next remark for an example).

Problem 4. If we consider a subset $Q \subset S$ of the state space S of a P system, can we find the invariant distribution restricted to Q?

Remark 4. This is probably *the most important* problem from a practical point of view. As already discussed in the first remark, when the P system is aperiodic and irreducible, the invariant distribution can be approximated by simulation. However, when the system leaves the subset Q during the simulations, knowledge about the dynamics outside of Q is needed, so this is not completely satisfying.

[3] Mathematically, although there does not exist a uniform probability distribution on V then, it is still possible to jump to a *random* element $j \in V$ with probability $\alpha > 0$ at each time step, when a probabilistic version of the *axiom of choice* is assumed. However, this will lead too far here, as it is not of practical importance.

The main problem with the calculation of the invariant distribution on Q is that the flow of probability from $S \setminus Q$ into Q is not known. However, for the invariant distribution the flow is in *equilibrium*, i.e., the total outflow from Q into its complement equals the total of the unknown inflows. By looping back each outflow into the inflows it is possible to constrain the possible invariant distributions[4] of Q, but it seems unlikely that they can be uniquely identified.

When inflows and matching outflows are *prescribed*, however, it is always possible to determine a corresponding invariant distribution on Q.

Problem 5. When is this invariant distribution compatible with prescribed outflows?

Remark 5. The answer is very simple: it never is, generically. So now it is important to find algorithmic ways of adjusting (and thereby violating) the inflow conditions such that the total sum of inflow and outflow violations is minimized.

Another related and very useful quantity, which *can* be computed by local information only is the mean *escape time*[5] from a subset $Q \subset S$ of state space.

4 Aperiodic and Irreducible P Systems

When a P system with probabilistic state transitions has an aperiodic and irreducible transition matrix, a unique invariant probability distribution exists. However, Google's random surfer model is not the only way to achieve these properties of the transition matrix.

For example, when all rules in a P system are irreversible, the system is irreducible. Unfortunately, in such a system the question of periodicity is unclear.

Let us now consider a more general situation. Assume that at each time step[6] there is a small probability for each object to change spontaneously into another object, analogous to mutations in DNA. The objects undergoing such a change

[4] Let Q be a finite set of order k. Consider the modified $(k+1)$-by-$(k+1)$ transition matrix Q_k that describes the transitions inside Q and from Q to an external state x (which represents all states outside of Q) and back from x into the k-th state of Q (with probability 1). Denote the unique invariant distribution of this matrix by r_i. Due to the linearity at the level of distributions, the true invariant distribution r, restricted to Q, is a linear combination of the first k components of all r_i. Of course, this only holds when $\alpha = 0$, but the result can easily be generalized to $\alpha > 0$ also.

[5] There are subtle connections between (i) this quantity, (ii) the page ranks of the states in Q, and (iii) the number of loops in Q. Basically, there are two contributions to the invariant distribution r on Q: Part of r consists of probability that flows into Q from outside of Q, and another part of r results from probability that flows around inside of Q in loops. However, this connection is not well understood at the moment, and further research is required.

[6] If time is assumed to be continuous, as in dynamical P systems that are simulated by Gillespie-type algorithms, there is still a discrete sequence of events, and by introducing an exponential waiting time distribution for such an event the same comments also apply to this case.

cannot be used in another rule at this time step. By adding rules of the form $u \to v$ for each possible pair of objects (u, v), we can realize this. Let us call a P system with this property a *leaky* P system. Note that leaky P systems are not always irreducible, as the example of a system with the rule $2A \to B$ shows: From the state with only one B we can never get to a state with more than one A, although the opposite is possible. However, if all rules were reversible, a leaky P system would be irreducible and have an invariant distribution.

A successful computation in a leaky P system would need to be robust against the continuous possibility of small changes of its objects. How could this be realized? What kind of error-correcting (repair) mechanisms can be envisaged in P systems? Moreover, the following two theoretical problems exist:

Problem 6. Given an irreducible aperiodic Markov chain, when adding the teleportation property of Eq. 1, do the corresponding invariant distributions $\mu(\alpha)$ converge in the limit $\alpha \to 0$?

Remark 6. Numerical studies have been done in case of the webgraph (confer [26] for references).

Problem 7. Although the random surfer model does not apply to a leaky P system, does the invariant distribution of a leaky P system converge against the same invariant distribution as in the random surfer model (of the same underlying P system), in the limit that $\alpha \to 0$?

Remark 7. This is the case if the leaky P system has the same communicating classes as the corresponding random surfer system P.

5 Recurrent Behavior and Complexity

A particular interesting application area for P systems is the emerging discipline of systems biology [24], and in the past years a number of biological systems have been simulated and analyzed by P systems [10,37]. It should be noted, however, that this line of research is only a small part of the total work on P systems, so we consider P systems from a particular perspective here.

The original state-transition P systems are characterized by a unique description of their dynamical behavior in terms of a *nondeterministic* and *maximally parallel* application of rules. The first of these concepts puts the focus not on an actual realization of behavior of a P system, but on all possible computations possible with it, i.e., on the (formal) *language* generated by it. The concept of maximal parallelism allows interesting control structures, but seems rather inappropriate when modeling in a biological context. Therefore, a number of researchers have turned to *dynamical* P system models, where the nondeterministic dynamics is replaced by a sequential and probabilistic evolution law. Two important approaches are *dynamically probabilistic P systems* [36] and the *metabolic algorithm* developed and propagated by V. Manca and colleagues [6,29]. The first is directly based on mass action kinetics [18], whereas the latter considers a special form of competition of rules for objects, called the *mass partition principle*. Another approach has been proposed in [37], where rules have fixed reaction rates.

As has been discussed in [31], static dynamical P systems are Markov chains.

Problem 8. How can notions from the theory of dynamical systems [14], such as fixed points and attractors be defined for dynamical P systems?

Remark 8. This problem underlies much of the recent research on dynamical P systems. Indeed, one has to be careful here. The notion of a dynamical system is (notwithstanding proper generalizations [3]) that of a deterministic system, whereas dynamical P systems are stochastic systems.

One consequence of this difference is that the notion of a fixed point, so useful in the theory of deterministic systems, has little importance in the theory of dynamical P systems. By definition the only fixed points in a dynamical P system are sinks, i.e., halting states.

We now give a satisfying solution of this problem. The proper way to analyze dynamical P systems from the dynamical perspective is by considering a proper generalization of fixed points, *recurrent behavior.* Different notions of recurrence are discussed in [1], the most general being *chain-recurrence*, introduced by Conley. A point x in a dynamical system is chain-recurrent, if for all $\epsilon > 0$ and $T > 0$, there exists a finite sequence of states $x = x_0, x_1, x_2, \ldots, x_n = x$ from x to itself, and a corresponding finite sequence of times $t_0, \ldots, t_{n-1}$ in $[T, \infty)$, such that the distance between x_{i+1} and the endpoint t_i of the trajectory, starting at x_i and being followed for a time t_i, is less than ϵ for all i. In other words, a chain-recurrent point can be reached by a sequence that alternately (i) follows the dynamics for at least a time T, and (ii) jumps to a state within a distance ϵ. It can be shown that the chain recurrent set contains all fixed points, periodic points and limit sets.

In a dynamical P system, the state space has the discrete topology, and time evolution is also discrete. A chain-recurrent point then corresponds to a point that is reachable from itself, i.e., the chain recurrent set is exactly the set of *communicating* states. So in an irreducible P system, which consists of exactly one communicating class, the chain recurrent set is the whole state space. But note that each dynamical P system, when started from a single initial condition (or a different initial condition that comes from the same communicating class) that lies inside a communicating class, is irreducible. What makes the notion of invariant distribution interesting, is that it carries quantitative information about how often the system is expected to be in a certain state. States with a higher page rank will be visited more often than states with a lower page rank. The invariant distribution *orders* the recurrent states of the system by their importance.

From a stationary distribution we can derive a complexity measure for P systems that quantifies the complexity in dynamical behavior.

Definition 1. *The* entropy *of a P system is the entropy of its invariant probability distribution. That is, if $p : S \mapsto \mathbb{R}^n_+$ is its invariant distribution, then*

$$h = \frac{-\sum_{i \in S} p(i) \log p(i)}{\log |S|}$$

is its entropy.

The denominator has been chosen such that $0 \leq h \leq 1$ holds. A low value of h signifies simple dynamical behavior, whereas a value of h close to one is characteristic of random behavior.

This idea generalizes the *global entropy* of [11], where a similar complexity measure has been introduced for probabilistic P systems with an evolution tree. Of course, the question arises what the advantage of such a measure is, compared to other complexity measures (for a list of possible candidates, see [9]). An important point here is that the definition of entropy of an invariant distribution is a mathematically elegant concept that quantifies the complexity of the dynamics of a P system in a way that relates to complexity considerations in other fields of science (confer [4,27]).

6 Approximating Asymptotic Behavior

Since the state space in dynamical P systems is usually infinite, a stationary distribution usually does not exist. Even if it does, it is not clear how to actually compute it. An interesting alternative is to simplify the situation considerably. Instead of working with the state space on which the dynamics takes place, we work with the *object network* of the P system, which is always finite.

Definition 2. *The* object network *of a P system is the directed graph $D = (V, A)$, where the vertex set V is given by the set of objects, and there exists an arc $(u, v) \in A$ between two objects $u, v \in V$ if there exists a rewriting rule of the form*

$$p_1 u_1 + \cdots + p_n u_n \to q_1 v_1 + \cdots + q_m v_m, \qquad p_k, q_l \geq 1 \textit{ for all } k \leq n, l \leq m,$$

and furthermore $u = u_i$ and $v = v_j$ for some indices $i \leq n$ and $j \leq m$.

The *connectivity matrix* of a P system is the the adjacency matrix of its object network, normalized row-wise such that its rows sum to one.

Definition 3. *The* ranking matrix *of a P system is the matrix $\bar{\bar{C}}$ (confer Eq. 1) where C is its connectivity matrix.*

Definition 4. *The* stationary object distribution *of a P system is the dominant eigenvector of its ranking matrix.*

The above definitions only use (i) the *topological* information about how objects can be transformed into each other. However, in a P system there are two more levels that can be considered, namely (ii) the stoichiometry, which introduces further constraints, and (iii) reaction rates. The latter has been discussed already, of course. However, incorporating the stoichiometry only, is not very satisfying. Eventually we need to come up with probabilities for a Markov chain, and although these can be readily defined from stoichiometric weights, this is a somewhat artificial construction that is difficult to interpret.

Let us finally, for completeness, consider the conventional analysis of steady state fluxes in biochemical networks [18]. Given a stoichiometric matrix $S \in$

$\mathbb{Z}^{m \times n}$ that describes the possible transitions of a chemical system, and some external fluxes $b \in \mathbb{R}_+^m$, one searches for a solution $x \in \mathbb{R}^n$ of the equation $S \cdot x = 0$, which is interpreted as a steady state flux. In the context of P systems, we can think of x as an *application vector*, telling us how often each rule has to be used. Unfortunately, linear algebra cannot be used, since the solutions need to be positive, i.e., it is necessary that $x_i \geq 0$ for some of the components of $x = (x_1, \ldots, x_n)$, since we cannot have negative rule applications. Therefore, one resorts to convex analysis and calculates the convex cone of all possible solutions [23]. This cone is usually not unique, so there are many possible steady state fluxes across the system.

But consider now what happens if we make use of the probabilities for transitions, corresponding to the complete probabilistic description of the system as in the beginning of the paper. The invariant distribution then induces a *unique* steady state flux (given by the product of the invariant distribution with the relative outdegrees), in contrast to the topological and the stoichiometric case. The implications of this, especially with regard to pathway analysis, have yet to be fully realized.

7 Page Ranking in P System Identification

In a previous work [30] we have discussed the general problem of identification of P systems; here we will focus on the application of page ranking to this problem. System identification can be considered the reverse of the usual modeling and analysis process. Instead of analyzing a *given* P system, the problem is to *find* an interesting P system that then can be analyzed, for example by simulation studies. This is particularly interesting in the application of P systems to biochemical systems. To this extent, public databases on the Internet can be used that store and collect information about biochemical reactions. These include WIT, EcoCyc, MetaCyc [22], aMAZE and KEGG [21]. For example, the LIGAND database [17], which is a particular database inside the KEGG repository, contains (as of version 42.0) information about 15053 chemical compounds (KEGG COMPOUND), 7522 biochemical reactions (KEGG REACTION) and 4975 enzymes (KEGG ENZYME) in ASCII text files that are easily parseable by computer.

In the usual approach [12,32] one constructs an *undirected* metabolite network graph $G = (V, E)$ from these files, where nodes represent compounds, and edges represent reactions (for simplicity, we do not consider enzymes here). Two compounds $u, v \in V$ in the metabolite graph are connected by an arc $(u, v) \in E \subseteq V^2$ if there exists a reaction in which both u and v participate. Note that u and v can both occur on the same side of a reaction, in contrast to what we have done for P system object networks, resulting in an undirected as opposed to a directed graph. The main problem considered in the bioinformatics community is the extraction of (meaningful) possible pathways that allow to transform one compound $s \in V$ into a target compound $t \in V$, which is equivalent to the k *shortest path problem* [13].

A particular problem with this approach is the existence of so-called *currency* metabolites [19]. These are usually small biomolecules that participate in a large number of reactions, and are used to store and transfer energy and/or certain ions. Examples of currency metabolites include H_2O, ATP, and NADH. Because of them, for example, there exist more than 500000 distinct pathways of length at most nine between glucose and pyruvate [25], most of which are not biochemically feasible. The solution considered by Croes and co-workers is to weight the paths by the (out-) degrees of their vertices, such that vertices with a large degree are punished relative to compounds with a higher specificity, i.e., a lower degree [12].

Here we propose to use a *directed* metabolite graph that more realistically captures the flow constraints of the biochemical reaction network, and to use the stationary distribution of such a biochemical object network to weight the paths. Currency metabolites are expected to have a large stationary probability, since they partake in many circular reaction patterns, and interesting pathways should then be found more effectively by bounding the total path weight.

P systems identification is then possible by first generating a large stoichiometric network graph, calculating its invariant distribution $p \in \mathbb{R}_+^N$, and using its components p_i, $1 \leq i \leq N$, to define weights $N \cdot p_i$ for a second pathway search (the constant N is used to ensure that the average weight is one). Only compounds encountered on paths with a weight below a certain, user-defined threshold are then used to define a P system model that captures the (hopefully) relevant biochemical reactions.

8 Discussion

In this paper we have shown some applications of page ranking to the analysis and identification of P systems. Dynamical P systems can be considered Markov chains, and Google's page ranking then corresponds to the stationary eigenvector of the transition matrix, after adding a small positive constant to ensure irreducibility and aperiodicity. For P systems, page ranking allows to define a probability distribution on the objects (and, dually, also on the rules), and this in turn allows to define the entropy of a P system, generalizing ideas of [11].

More generally, this work was motivated by the urge to adapt the methods of dynamical systems theory to P systems, and from this perspective the invariant distribution of a P system can be considered to represent the recurrent dynamical behavior. In particular, we can now give operational definitions of the concept of "fixed points" for P systems as states with large invariant probability, whereas "transient" states will have very small invariant probability (on the order of α).

A different application has been in the identification of P system models from biochemical databases. The invariant distribution should allow to search more effectively for pathways, improving the degree weights introduced by Croes and co-workers. Although the complete stoichiometric graph available in the LIGAND database consists of more than 10000 vertices, the eigenvector calculation has to be done only once. The test of this idea is underway.

Acknowledgements. This work has been supported by the Dutch Science Foundation (NWO) under grant no. 635.100.006.

References

1. Alongi, J.M., Nelson, G.S.: Recurrence and Topology. American Mathematical Society (2007)
2. Altman, A., Tennenholtz, M.: Ranking systems: the PageRank axioms. In: Proc. 6th ACM conference on Electronic Commerce, pp. 1–8 (2005)
3. Arnold, L.: Random Dynamical Systems. Springer, Heidelberg (1998)
4. Badii, R., Politi, A.: Complexity. Hierarchical Structures and Scaling in Physics. Cambridge University Press, Cambridge (1997)
5. Bapat, R.B., Raghavan, T.E.S.: Nonnegative Matrices and Applications. Cambridge University Press, Cambridge (1997)
6. Bianco, L., Fontana, F., Manca, V.: P systems with reaction maps. Int. J. Found. Comp. Sci. 17, 3–26 (2006)
7. Brémaud, P.: Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues. Springer, Heidelberg (1999)
8. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30, 107–117 (1998)
9. Chakrabarti, D., Faloutsos, C.: Graph mining: Laws, generators, and algorithms. ACM Computing Surveys 38, 1–69 (2006)
10. Ciobanu, G., Păun, G., Pérez-Jiménez, M.J. (eds.): Applications of Membrane Computing. Springer, Heidelberg (2006)
11. Cordón-Franco, A., Sancho-Caparrini, F.: A note on complexity measures for probabilistic P systems. J. Universal Computer Sci. 10, 559–568 (2004)
12. Croes, D., Couche, F., Wodak, S.J., van Helden, J.: Inferring meaningful pathways in weighted metabolic networks. J. Mol. Bio. 356, 222–236 (2006)
13. Epstein, E.: Finding the k shortest paths. SIAM J. Comput. 28, 652–673 (1998)
14. Guckenheimer, J., Holmes, P.: Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields. Springer, Heidelberg (1983)
15. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press (1996)
16. Google: Google Technology, `http://www.google.com/technology/`
17. Goto, S., Nishioka, T., Kanehisa, M.: LIGAND: chemical database for enzyme reactions. Bioinformatics 14, 591–599 (1998)
18. Heinrich, R., Schuster, S.: The Regulation of Cellular Systems. Springer, Heidelberg (1996)
19. Huss, M., Holme, P.: Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. IET Syst. Biol. 1, 280–285 (2007)
20. Kamvar, S., Haveliwala, T., Golub, G.: Adaptive methods for the computation of PageRank. Linear Algebra Appl. 386, 51–65 (2004)
21. Kanehisa, M., Araki, M., Goto, S., Hattori, M., et al.: KEGG for linking genomes to life and the environment. Nucleic Acids Research 484, D380–D484.(2008)
22. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., et al.: The EcoCyc and MetaCyc databases. Nucleic Acids Research 28, 56–59 (2000)
23. Kauffman, K.J., Prakash, P., Edwards, J.S.: Advances in flux balance analysis. Current Opinion in Biotechnology 14, 491–496 (2003)
24. Klipp, E., Herwig, R., Kowald, A., Wierling, C., et al.: Systems Biology in Practice. Wiley-VCH, Chichester (2005)
25. Kuffner, R., Zimmer, R., Lengauer, T.: Pathway analysis in metabolic databases via differential metabolic display (DMD). Bioinformatics 16, 825–836 (2000)

26. Langville, A.N., Meyer, C.D.: Deeper inside PageRank. Internet Math. 1, 335–380 (2004)
27. Lind, D., Marcus, B.: An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge (1995)
28. Ma, N., Guan, J., Zhao, Y.: Bringing PageRank to the citation analysis. Information Processing & Management 44, 800–810 (2008)
29. Manca, V., Bianco, L.: Biological networks in metabolic P systems. BioSystems 91, 489–498 (2008)
30. Muskulus, M.: Identification of P system models assisted by biochemical databases. In: Ibarra, O.H., Sosík, P. (eds.) Prague International Workshop on Membrane Computing, Preliminary Proc., pp. 46–49 (2008)
31. Muskulus, M., Besozzi, D., Brijder, R., Cazzaniga, P., et al.: Cycles and communicating classes in membrane systems and molecular dynamics. Theoretial Computer Sci. 372, 242–266 (2007)
32. Noirel, J., Ow, S.Y., Sanguinetti, G., Jaramillo, A., et al.: Automated extraction of meaningful pathways from quantitative proteomics data. Briefings in Functional Genomics and Proteomics (in press)
33. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking. Bringing order to the web. Technical Report, Stanford University (1998), `http://dbpubs.stanford.edu:8090/pub/1999-66`
34. Păun, G.: Computing with membranes. J. Computer System Sci. 61, 108–143 (2000)
35. Păun, G.: Membrane Computing. An Introduction. Springer, Heidelberg (2002)
36. Pescini, D., Besozzi, D., Mauri, G., Zandron, C.: Dynamical probabilistic P systems. Intern. J. Found. Comp. Sci. 17, 183–204 (2006)
37. Romero-Campero, F.J., Pérez-Jiménez, M.J.: Modelling gene expression control using P systems. The Lac operon, a case study. Biosystems 91, 438–457 (2008)
38. Vise, D., Malseed, M.: The Google Story. Random House (2006)