Proper down-coloring simple acyclic digraphs

Geir Agnarsson[1], Ágúst S. Egilsson[2], and Magnús M. Halldórsson[3]

[1] Department of Mathematical Sciences, George Mason University,
MS 3F2, 4400 University Drive, Fairfax, VA 22030,
geir@math.gmu.edu
[2] Department of Mathematics, University of California,
Berkeley, CA 94720-3840,
egilsson@Math.Berkeley.EDU
[3] Department of Computer Science, University of Iceland,
Dunhaga 3, IS-107 Rvk, Iceland,
mmh@hi.is

Abstract. We consider vertex coloring of a simple acyclic digraph $\overline{G}$ in such a way that two vertices which have a common ancestor in $\overline{G}$ receive distinct colors. Such colorings arise in a natural way when clustering, indexing and bounding space for various genetic data for efficient analysis. We discuss the corresponding chromatic number and derive an upper bound as a function of the maximum number of descendants of a given vertex and the inductiveness of the corresponding hypergraph, which is obtained from the original digraph. Finally we obtain some general approximation results.

Keywords: Relational databases, multidimensional clustering, indexing, bitmap indexing, vertex coloring, digraph, poset, hypergraph, ancestor, descendant, down-set.

1 Introduction

The purpose of this article is to discuss a special kind of vertex coloring for acyclic digraphs in bioinformatics, where we insist upon two vertices, that have a common ancestor, receive distinct colors.

This article can be viewed as a continuation of [1], but we have attempted to make it as self-contained as possible. For a brief introduction and additional references to the ones mention here, we refer to [1].

Before setting forth our terminology and stating precise definitions we need to address the justification for such vertex colorings and why they are of interest, especially for various databases containing life science related data.

Digraphs representing various biological phenomena and knowledge are used throughout in the life sciences and in drug discovery research, i.e. the gene ontology digraph maintained by the Gene Ontology Consortium, [2]. Many other open biological ontologies are available, and additionally medical or disease classifications in the form of acyclic digraphs are used throughout. In a typical setting the digraphs are referenced by numerous large tables containing genetic or other observed data and stored in relational data warehouses. An overview of several projects relating to indexing of semistructured data (i.e. graphs) can be found in [3].

The challenges addressed in this article include certain coloring analysis of an acyclic digraph $\overline{G}$, which provides an efficient structure for querying of relational tables referencing the digraph $\overline{G}$. This includes efficiently identifying and retrieving all rows in a given table that are conditioned, based on sets of ancestors of $\overline{G}$.

Consistent with the notation introduced in the next section, we let $U[u] = \{u\} \cup \{x : x \text{ is an ancestor of } u\}$ for each vertex u of a a given digraph $\overline{G}$. Also, assign a unique data entry $U[u]*$ to each of the ancestor sets considered. Since each ancestor set $U[u]$ is determined uniquely by the vertex u, then we can simply identify u and $U[u]*$, and assume that $u = U[u]*$.

In order to create an index on a table column referencing the vertices $V(\overline{G})$ of the digraph $\overline{G}$, it is necessary to develop a schema, such as a collection of functions, for assigning data entries to the search key values in the table. First though, we develop such a schema for the vertices $V(\overline{G})$ themselves. For a given vertex u there should be at least one function f, in the schema, defined on a subset of $V(\overline{G})$, satisfying $U[u] = f^{-1}(U[u]*)$, that is, f assigns the data entry $U[u]*$ only to search key elements from the set $U[u]$. Therefore, a complete schema is a collection $f_1, f_2, \ldots, f_k$ of functions so that for each vertex u there exists at least one integer $c(u)$ with

$$U[u] = f_{c(u)}^{-1}(U[u]*).$$

The vertex coloring $u \mapsto c(u)$ is what we will call a *down-coloring* of the digraph $\overline{G}$, as we will define in the next section, and it has the the following property: If two distinct vertices u and v have a common ancestor, say w, then $f_{c(u)}(w) = U[u]*$ and $f_{c(v)}(w) = U[v]*$, so since $U[u]* \neq U[v]*$, we must have $c(u) \neq c(v)$, that is u and v receive distinct colors. This allows us to conclude that k is at least the *down-chromatic number* of $\overline{G}$, defined in the next section.

Conversely, given a down-coloring $c : V(\overline{G}) \to \{1, \ldots, k\}$ one can construct a complete schema for assigning data entries to the vertices by defining functions $f_1, \ldots, f_k$ as

$$f_i(u) = U[v]*, \text{ if } u \in U[v] \text{ and } c(v) = i.$$

The down-coloring condition ensures that the functions are well defined. The possible schemas of “functional indexes” are therefore in a one-to-one correspondence with the possible down-colorings.

One can realize the schema $f_1, \ldots, f_k$ in a relational database system in many different ways. In the relational database system, one may try additionally to devise a structure so that the set operations ($\cup$, $\cap$, $\setminus$) can be optimally executed on elements from the collection $\{U[u] : u \in V(\overline{G})\}$. A straightforward way to do this is to have the functions physically share the domain $V(\overline{G})$ in the database, instead of using additional equijoins to implement the sharing of the domain. For digraphs with relatively small down-chromatic numbers we therefore materialize the relation

$$\{(u, f_1(u), \ldots, f_k(u)) : u \in V(\overline{G})\}$$

in the database system, in addition to the coloring map c. This requires that $f_j(u) =$ "NULL" if u is not in the domain of f_j, following standard convention. The table containing the above relation is referred to as Clique(U), it has a domain column "Vertex" and a column "CJ" for each of the colors $j \in \{1, 2, \ldots, k\}$. Of course, it is also the graph of the function $f_1 \times f_2 \times \cdots \times f_k$.

If a (large) table references the digraph $\overline{G}$ in one of its columns, then there are several possible methods to index the column using the data entries schema Clique(U) or, equivalently, the functions $f_1, \ldots, f_k$. Below we summarize two of the possible methods.

1. The Clique(U) relation may be joined (through the "Vertex" column) with any table that references the digraph $\overline{G}$ in one of its columns. In this way the Clique(U) table may be viewed as a dimension table, allowing queries conditioned on the ancestors sets and other conditioning to be evaluate as star queries. Additional search mechanisms are introduced, such as bitmap and other indexes and, most efficiently, bitmap join indexes may be used (e.g., one for each color column in Clique(U) when using the Oracle 9i system). Both the Oracle database system and DB2 from IBM are able to take advantage of this simple and space efficient setup for evaluating queries, see [4] and [5] for further information.
2. A table referencing the digraph $\overline{G}$ in one of its columns may be joined with, and clustered according to the schema Clique(U) using multidimensional clustering, if the down-chromatic number for the digraph is small. Ideally, the multidimensional clustering enables the referencing table to be physically clustered along all the color columns simultaneously. Commercially available, relational database systems, that implement multidimensional clustering, include IBM's DB2 Universal Database (version 8.1), see [6] for an introduction to the feature in DB2.

The first method is currently used by deCODE Genetics in Iceland. There the Clique(U) structure and bitmap and sometimes bitmap join indexes are combined to map a data entry $U[u]*$ to all the rows in a table referencing elements from $U[u]$. Comparisons, in this setting, favor greatly using the Clique(U) structure and star-query methods over other methods. We will demonstrate this with an example:

A small table with about 1.5 million rows called "goTermFact" references the gene ontology digraph in a column called "acc". Additionally, the table has several other columns including a number column denoted by "m". The acyclic gene ontology digraph has 14,513 edges and 11,368 vertices, the edges are directed and converge at a root vertex, the root has three ancestors called "molecular function", "biological process" and "cellular component". One of the ancestors of "molecular function" in the graph is called "enzyme activity". We wish to summarize the column "m" for all the 394,702 rows that reference an ancestor of the vertex "enzyme activity". The digraph is down-colored using 36 colors and the vertex “enzyme activity” receives color “8”, it is also assigned code “GO:0003824” by the Gene Ontology Consortium. The SQL query (Q1) is constructed as:

```
Q1 : select sum(f.m) from goTermFact f, clique_U d where
          f.acc = d.vertex and d.C8 = ‘GO:0003824’
```

For the purpose of comparison, the referencing table can also be indexed using the digraph, in a more traditional way, by creating a lookup-table “Lookup” with columns “path” and “rid”. The “path” column is a string path starting at the root vertex and the “rid” column contains row-ids from the referencing table. Since this particular digraph is not a tree structure, then one possibly needs to store several paths for each vertex (1.9 on the average in this case). The “Lookup” table is stored as an index-organized table with a composite primary key (“path” and “rid”) in the Oracle 9.2i database, providing fast primary key based access to lookup data for range search queries. A path from the root vertex to the “enzyme activity” vertex, following the (reversed) digraph, is of the form: ‘GO:0003673.GO:0003674.GO:0003824’. Therefore the above query (Q1) can now be rewritten using a range search of the “Lookup” table as follows:

```
Q2 : select sum(m) from goTermFact where
          rowid in (select rid from Lookup where
            path like ‘GO:0003673.GO:0003674.GO:0003824%’)
```

Executing the queries on a Windows XP based Pentium III, 1.2GHz, 512MB system using the Oracle 9.2i database reveals the following comparison:

Q1: Form Q1 executes up to 150 times faster than equivalent form Q2. The best performance is achieved using a bitmap join index to (pre-)resolve the join between the Clique(U) relation and the table. The query takes between 0.03 sec and 21 sec, depending on whether the data is located in memory or on disk. The query is also efficiently executed by using a (static) bitmap index on the “acc” column and the bitmap OR operator to dynamically construct, using Clique(U), a pointer to all the rows in the “goTermFact” table that satisfy the join.

Q2: In form Q2 the query takes between 4.02 sec and 104 sec, depending again on whether the table and index data is located in memory or on disk when the query is executed. Clearly, this form is much more space consuming and therefore less efficient then the previous one.

The above comparison, as well as many other results obtained using similar clique indexing schemas, demonstrate the power of the indexing, when combined with current relational database optimization techniques.

Directed acyclic graphs are many times called DAG’s by computer scientists, as is the case in [7, p. 194], but we will here call them acyclic digraphs.

Only few vertex colorings results rely on the structure that the direction of the edges in a digraph provide. In [8] the *arc-chromatic number* for a digraph $\overline{G}$ is investigated, and in [9] and [10] the *dichromatic number* of a digraph is studied.

Here in this article we define the *down-chromatic number* of an acyclic digraph, discuss some of its properties, similarity and differences with ordinary graph coloring, and derive an upper bound which, in addition, yields an efficient coloring procedure. We will give some different representations of our acyclic digraph, some equivalent and other more relaxed representations, which, from our vertex coloring point of view, will suffice to consider.

2 Definitions and observations

We attempt to be consistent with standard graph theory notation in [11], and the notation in [12] when applicable.

For a natural number $n \in \mathbb{N}$ we let $[n] = \{1, \ldots, n\}$. A *simple digraph* is a finite simple directed graph $\overline{G} = (V, E)$, where V is a finite set of vertices and $E \subseteq V \times V$ is a set of directed edges. For a digraph $\overline{G}$, the sets of its vertices and edges will many times be given by $V(\overline{G})$ and $E(\overline{G})$ respectively. The digraph $\overline{G}$ is said to be *acyclic* if $\overline{G}$ has no directed cycles. Henceforth $\overline{G}$ will denote an acyclic digraph in this section.

The binary relation $\leq$ on $V(\overline{G})$ defined by

$$u \leq v \Leftrightarrow u = v, \text{ or there is a directed path from } v \text{ to } u \text{ in } \overline{G}, \qquad (1)$$

is a reflexive, antisymmetric and transitive binary relation, and therefore a partial order on $V(\overline{G})$. Hence, whenever we talk about $\overline{G}$ as a *poset*, the partial order will be the one defined by (1). Note that the acyclicity of $\overline{G}$ is essential in order to be able to view $\overline{G}$ as a poset. By the *height* of $\overline{G}$ as a poset, we mean the number of vertices in the longest directed path in $\overline{G}$. We denote by $\max\{\overline{G}\}$ the maximal vertices of $\overline{G}$ with respect to the partial order $\leq$.

For vertices $u, v \in V(\overline{G})$ with $u \leq v$, we say that u is a *descendant* of v, and v is an *ancestor* or u. The *open principal down-set* $D(u)$ of a vertex $u \in V(\overline{G})$ is the set of all descendants of u in $\overline{G}$, that is, $D(u) = \{x \in V(\overline{G}) : x < u\}$. Similarly the *closed principal down-set* $D[u]$ of a vertex $u \in V(\overline{G})$ is the set of all descendants of u in $\overline{G}$, including the vertex u itself, that is, $D[u] = \{x \in V(\overline{G}) : x \leq u\} = D(u) \cup \{u\}$.

Definition 1. *A* down-coloring *of $\overline{G}$ is a map $c : V(\overline{G}) \to [k]$ satisfying*

$$u, v \in D[w] \textit{ for some } w \in V(\overline{G}) \ \Rightarrow c(u) \neq c(v)$$

for every $u, v \in V(\overline{G})$. The down-chromatic *number of $\overline{G}$, denoted by $\chi_d(\overline{G})$, is the least k for which $\overline{G}$ has a proper down-coloring $c : V(\overline{G}) \to [k]$.*

Just as in an undirected graph G, the vertices in a clique must all receive distinct colors in a proper vertex coloring of G. Therefore $\omega(G) \leq \chi(G) \leq |V(G)|$ where $\omega(G)$ denotes the clique number of G. In addition $\chi(G)$ can be larger than the clique number $\omega(G)$. Similarly for our acyclic digraph $\overline{G}$ we have $D(\overline{G}) \leq \chi_d(\overline{G}) \leq |V(\overline{G})|$, where

$$D(\overline{G}) = \max_{u \in V(\overline{G})} \{|D[u]|\},$$

and we will see below that the same holds for $\overline{G}$, that $\chi_d(\overline{G})$ can be much larger than $D(\overline{G})$.

A useful approach to consider down-colorings of a given acyclic digraph $\overline{G}$, is to construct the corresponding simple undirected *down-graph* G' on the same set of vertices as $\overline{G}$, but where we connect each pair of vertices that are contained in the same principal down-set

$$\begin{aligned} V(G') &= V(\overline{G}), \\ E(G') &= \{\{u,v\} : u, v \in D[w] \text{ for some } w \in V(\overline{G})\}. \end{aligned}$$

In this way we have transformed the problem of down-coloring the digraph $\overline{G}$ to the problem of vertex coloring the simple undirected graph G' in the usual sense, and we have $\chi_d(\overline{G}) = \chi(G')$. Hence, from the point of down-colorings, both $\overline{G}$ and G' are equivalent, which is something we will discuss in the next section to come. However, some structure is lost. The fact that two vertices u and v are connected in G' could mean one of three possibilities, $u < v$, $u > v$, or u and v are incomparable, but there is a vertex w with $u < w$ and $v < w$.

Although a down-set in $\overline{G}$ will become a clique in G', the converse is not true, as stated in Observation 1 below.

We conclude this section with a concrete example and some observations drawn from it. A special case of this following example can be found in [1].

EXAMPLE: Let $k \geq 2$ and $m \geq 1$ be natural numbers. Let $A_1, \ldots, A_k$ be disjoint sets, each A_i containing exactly m vertices. For each of the $\binom{k}{2}$ pairs $\{i,j\} \subseteq [k]$ define an additional vertex w_{ij}. Let $\overline{G}(k,m)$ be a digraph with vertex set and edge set given by

$$\begin{aligned} V(\overline{G}(k,m)) &= \left(\bigcup_{i\in[k]} A_i\right) \cup \{w_{ij} : \{i,j\} \subseteq [k]\}, \\ E(\overline{G}(k,m)) &= \bigcup_{\{i,j\}\subseteq[k]} \{(w_{ij}, u) : u \in A_i \cup A_j\}. \end{aligned}$$

Clearly $\overline{G}(k,m)$ is a simple acyclic digraph on $n = km + \binom{k}{2}$ vertices and with $\binom{k}{2} \cdot 2m = k(k-1)m$ directed edges. Each closed principal down-set is of the form $D[w_{ij}]$ for some $\{i,j\} \subseteq [k]$ and hence $D(\overline{G}(k,m)) = 2m + 1$. Note that in any proper down-coloring of $\overline{G}(k,m)$, every two vertices in $\bigcup_{i\in[k]} A_i$ are both contained in $D[w_{ij}]$ for some $\{i,j\} \subseteq [k]$, and hence $\bigcup_{i\in[k]} A_i$ forms a clique in $G'(k,m)$. From this we see that $\omega(G'(k,m)) = km$. In particular we have that

$$\chi_d(\overline{G}(k,m)) = \chi(G'(k,m)) \geq \omega(G'(k,m)) = km.$$

In fact, any coloring of $\bigcup_{i\in[k]} A_i$ with km colors can be extended in a greedy fashion to a proper down-coloring of $\overline{G}(k,m)$ with at most km colors. Therefore equality holds through the above display. In particular we have $D(\overline{G}(k,1)) = 3$ and $\omega(G'(k,1)) = k$, from which we deduce the following observation.

Observation 1 *There is no function $f : \mathbb{N} \to \mathbb{N}$ with $\omega(G') \leq f(D(\overline{G}))$ for every simple acyclic digraph $\overline{G}$. In particular, there is no function f with $\chi_d(\overline{G}) \leq f(D(\overline{G}))$ for all simple acyclic digraphs $\overline{G}$.*

Let $\alpha \in \mathbb{N}$ be a fixed and "large" natural number. Denoting by n the number of vertices of $\overline{G}(k, \alpha k)$, we clearly have $n = |V(\overline{G}(k, \alpha k))| = \alpha k^2 + \binom{k}{2} \sim k^2(\alpha + 1/2)$, where $f(k) \sim g(k)$ means $\lim_{k\to\infty} f(k)/g(k) = 1$. We now see that

$$D(\overline{G}(k, \alpha k)) = 2\alpha k + 1 \sim \left(\frac{2\alpha}{\sqrt{\alpha + 1/2}}\right)\sqrt{n},$$
$$\chi_d(\overline{G}(k, \alpha k)) = \alpha k^2 \sim \left(\frac{\alpha}{\alpha + 1/2}\right) n.$$

From this we have the following.

Observation 2 *For every $\epsilon > 0$, there is an $n \in \mathbb{N}$ for which there is a simple acyclic digraph $\overline{G}$ on n vertices with $D(\overline{G}) = \Theta(\sqrt{n})$ and $\chi_d(\overline{G}) \geq (1 - \epsilon)n$.*

REMARK: The above Observation 2 simply states that $\chi_d(\overline{G})$ can be an arbitrarily large fraction of $|V(\overline{G})|$ without making $D(\overline{G})$ too large. Hence, both the upper and lower bound in the inequality $D(\overline{G}) \leq \chi_d(\overline{G}) \leq |V(\overline{G})|$ are tight in this sense.

3 Various representations

In this section we discuss some different representation of our digraph $\overline{G}$, and define some parameters which we will use to bound the down-chromatic number $\chi_d(\overline{G})$.

We first consider the issue of the height of digraphs. We say that two digraphs on the same set of vertices, are *equivalent* if every down-coloring of one is also a valid down-coloring of the other, that is, if they induce the same undirected down-graph. We show that for any acyclic digraph $\overline{G}$ there is an equivalent acyclic digraph $\overline{G}_2$ of height two with $\chi_d(\overline{G}) = \chi_d(\overline{G}_2)$. However, the degrees of vertices in $\overline{G}_2$ may necessarily be larger than in $\overline{G}$.

Proposition 1. *Any down-graph G' of an acyclic digraph $\overline{G}$ is also a down-graph of an acyclic digraph $\overline{G}_2$ of height two.*

Proof. The derived digraph $\overline{G}_2$ has the same vertex set as $\overline{G}$, while the edges all go from $\max\{\overline{G}\}$ to $V(\overline{G}) \setminus \max\{\overline{G}\}$, where $(u, v) \in E(\overline{G}_2)$ if, and only if, $v \in D[u]$. In this way we see that two vertices in $\overline{G}$ have a common ancestor if, and only if they have a common ancestor in $\overline{G}_2$. Hence, we have the proposition.

Therefore, when considering down-colorings of digraphs, we can by Proposition 1 assume them to be of height two. Moreover, there is a natural correspondence between acyclic digraphs and certain hypergraphs.

Definition 2. *For a digraph $\overline{G}$, the related* down-hypergraph $H_{\overline{G}}$ *of $\overline{G}$ is given by:*

$$V(H_{\overline{G}}) = V(\overline{G})$$
$$E(H_{\overline{G}}) = \{D[u] : u \in \max\{\overline{G}\}\}.$$

Note that the down-graph G' is the *clique graph* of the down-hypergraph $H_{\overline{G}}$, that is, the simple undirected graph where every pair of vertices which are contained in a common hyperedge in $H_{\overline{G}}$ are connected by an edge in G'.

As we shall see, not every hypergraph is a down-hypergraph. There is a simple criteria for whether a hypergraph is a down-hypergraph or not. An edge in a hypergraph has a *unique-element* if it contains a vertex contained in no other edge. A hypergraph has the *unique-element property* if every edge has a unique element.

Observation 3 *A hypergraph is a down-hypergraph if, and only if, it has the unique-element property.*

Proof. A down-hypergraph is defined to contain the principal closed down-sets of a digraph $\overline{G}$ as edges. Each such edge contains a maximal element in $\overline{G}$, and this element is not contained in any other down-set. Hence, a down-hypergraph satisfies the unique-element property.

Suppose a hypergraph H satisfies the property. Then we can form a height-two acyclic digraph as follows: For each hyperedge, add a source vertex in the digraph corresponding to the representative unique element of the hyperedge. For the other hypervertices, add sinks to the digraph with edges from the sources to those sinks that correspond to vertices in the same hyperedge.

Note that a hypergraph with the unique element property is necessarily *simple*, in the sense that each hyperedge is uniquely determined by the vertices it contains.

We see that we can convert a proper down-hypergraph to a corresponding acyclic digraph of height two, and vice versa, in polynomial time.

Definition 3. *A* strong coloring *of a hypergraph H, is a map $\Psi : V(H) \to [k]$ satisfying*

$$u, v \in e \text{ for some } e \in E(H) \Rightarrow \Psi(u) \neq \Psi(v).$$

The strong chromatic number $\chi_s(H)$ *is the least number k of colors for which H has a proper strong coloring $\Psi : V(H) \to [k]$.*

Note that a strong coloring of a down-hypergraph $H_{\overline{G}}$ is equivalent to a down-coloring of $\overline{G}$, and hence $\chi_s(H_{\overline{G}}) = \chi_d(\overline{G})$. Since we can convert to and from hypergraph and digraph representations, the two coloring problems are polynomial-time reducible to each other. Strong colorings of hypergraphs have been studied, but not to the extent of various other types of colorings of hypergraphs. In [13] a nice survey of various aspects of hypergraph coloring theory is found, containing almost all fundamental results in the past three decades.

In the next section we will bound the down-chromatic number of our acyclic digraph $\overline{G}$, partly by another parameter of the corresponding down-hypergraph $H_{\overline{G}}$.

4 Upper bound in terms of degeneracy

As we saw in Observation 1, it is in general impossible to bound $\chi_d(\overline{G})$ from above solely in terms of $D(\overline{G})$, even when $\overline{G}$ is of height two. Therefore we need an additional parameter for that upper bound, but first we need to review some matters about a hypergraph $H = (V(H), \mathcal{E}(H))$.

Two vertices in $V(H)$ are *neighbors* in H if they are contained in the same edge in $\mathcal{E}(H)$. An edge in $\mathcal{E}(H)$ containing just one element is called *trivial.* The largest cardinality of a hyperedge of H will be denoted by $\sigma(H)$. The *degree* $d_H(u)$, or just $d(u)$, of a vertex $u \in V(H)$ is the number of non-trivial edges containing u. Note that $d(u)$ is generally much smaller than the number of the neighbors of u. The minimum and maximum degree of H are given by

$$\begin{aligned}\delta(H) &= \min_{u \in V(H)} \{d_H(u)\},\\ \Delta(H) &= \max_{u \in V(H)} \{d_H(u)\}.\end{aligned}$$

The subhypergraph $H[S]$ of H, induced by a set S of vertices, is given by

$$\begin{aligned}&V(H[S]) = S.\\ &\mathcal{E}(H[S]) = \{X \cap S : X \in \mathcal{E}(H) \text{ and } |X \cap S| \geq 2\}.\end{aligned}$$

Definition 4. *Let H be a simple hypergraph. The* degeneracy *or the* inductiveness *of H, denoted by* ind(H)*, is given by*

$$\mathrm{ind}(H) = \max_{S \subseteq V(H)} \{\delta(H[S])\}.$$

*If $k \leq \mathrm{ind}(H)$, then we say that H is k-*degenerate *or k-*inductive.

Note that Definition 4 is a natural generalization of the degeneracy or the inductiveness of a usual undirected graph G, given by $\mathrm{ind}(G) = \max_{H \subseteq G} \{\delta(H)\}$, where H runs through all the induced subgraphs of G. Note that the inductiveness of a (hyper)graph is always greater than or equal to the inductiveness of any of its sub(hyper)graph.

To illustrate, let us for a brief moment discuss the degeneracy of an important class of simple graphs, namely that of simple planar graphs. Every subgraph of a simple planar graph is again planar. Since every planar graph has a vertex of degree five or less, the degeneracy of every planar graph is at most five. This is the best possible for planar graphs, since the graph of the icosahedron is planar and 5-regular. That a planar graph has degeneracy of five, implies that it can be vertex colored in a simple greedy fashion with at most six colors. The degeneracy has also been used to bound the chromatic number of the square G^2 of a planar graph G, where G^2 is a graph obtained from G by connecting two vertices of G if, and only if, they are connected in G or they have a common neighbor in G, [14].

In general, the degeneracy of an undirected graph G yields an ordering $\{u_1, u_2, \ldots, u_n\}$ of $V(G)$, such that each vertex u_i has at most ind(G) neighbors among the previously listed vertices $u_1, \ldots, u_{i-1}$. Such an ordering provides a way to vertex color G with at most $\mathrm{ind}(G)+1$ colors in an efficient greedy way, and hence we have in general that $\chi(G) \leq \mathrm{ind}(G)+1$.

The inductiveness of a simple hypergraph is also connected to a greedy vertex coloring of it, but not in such a direct manner as for regular a undirected graph, since, as noted, the number of neighbors of a given vertex in a hypergraph is generally much larger than its degree.

Theorem 4. *If the simple undirected graph G is the clique graph of the simple hypergraph H then* $\mathrm{ind}(G) \leq \mathrm{ind}(H)(\sigma(H)-1)$.

Proof. For each $S \subseteq V(G) = V(H)$, let $G[S]$ and $H[S]$ be the subgraph of G and the subhypergraph of H induced by S, respectively. Note that for each $u \in S$, each hyperedge in $H[S]$ which contains u, has at most $\sigma(H[S]) - 1 \leq \sigma(H) - 1$ other vertices in addition to u. By definition of $d_{H[S]}(u)$, we therefore have that $d_{G[S]}(u) \leq d_{H[S]}(u)(\sigma(H)-1)$, and hence

$$\delta(G[S]) \leq \delta(H[S])(\sigma(H)-1). \tag{2}$$

Taking the maximum of (2) among all $S \subseteq V(G)$ yields the theorem.

Observation 5 *For a simple connected hypergraph H, then* $\mathrm{ind}(H) = 1$ *if, and only if, the intersection graph of its hyperedges $\mathcal{E}(H)$ is a tree.*

What Observation 5 implies, is that edges of H can be ordered as $\mathcal{E}(H) = \{e_1, \ldots, e_m\}$, such that each e_i intersects exactly one edge from $\{e_1, \ldots, e_{i-1}\}$. If now G is the clique graph of H, this implies that $\mathrm{ind}(G) = \sigma(H) - 1$.

Also note that if H has the unique element property and $\sigma(H) = 2$, then clearly the clique graph G is a tree, and hence $\mathrm{ind}(G) = 1 = \sigma(H) - 1$. We summarize in the following.

Observation 6 *Let H be a hypergraph H that satisfies the unique element property. If either* $\mathrm{ind}(H) = 1$ *or* $\sigma(H) = 2$*, then the clique graph G of H satisfies* $\mathrm{ind}(G) = \sigma(H) - 1$.

For a hypergraph H with the unique element property, we can obtain some slight improvements in the general case as well.

Theorem 7. *Let H be a hypergraph with the unique element property. Assume further that* $\mathrm{ind}(H) > 1$ *and* $\sigma(H) > 2$*. Then the graph G of H satisfies*

$$\mathrm{ind}(G) \leq \mathrm{ind}(H)(\sigma(H)-2).$$

Proof. Since H has the unique element property, then by Observation 3 there is an acyclic digraph $\overline{G}$ such that $H = H_{\overline{G}}$. Let H'' be the hypergraph induced by $V(H) \setminus \max\{\overline{G}\}$ and G'' be the corresponding clique graph of H''. Since each $u \in \max\{\overline{G}\}$ is simplicial in H and in G, their removal will not effect the degeneracy of the remaining vertices, so $\mathrm{ind}(H'') = \mathrm{ind}(H)$ and $\mathrm{ind}(G'') = \mathrm{ind}(G)$. Also note that $\sigma(G'') = \sigma(G) - 1$. By Theorem 4 we get $\mathrm{ind}(G) = \mathrm{ind}(G'') \leq \mathrm{ind}(H'')(\sigma(H'')-1) = \mathrm{ind}(H)(\sigma(H)-1)$, thereby completing the proof.

REMARK: Recall that for any simple undirected graph G on n vertices, we have $\chi(G) \leq \operatorname{ind}(G)+1$. In fact, the upper bounds of $\operatorname{ind}(G)$ given in Theorem 7 yields an online down-coloring of $\overline{G}$ that uses at most $\operatorname{ind}(G)\log n$ colors, where $H = H_{\overline{G}}$ is the down-hypergraph of $\overline{G}$, as well.

Let $\overline{G}$ be an acyclic digraph. Since now $D(\overline{G}) = \sigma(H_{\overline{G}})$ and $\chi_d(\overline{G}) = \chi(G') \leq \operatorname{ind}(G')+1$, we obtain the following summarizing corollary.

Corollary 1. *If $\overline{G}$ is an acyclic digraph, then its down-chromatic number satisfies the following:*

1. *If $\operatorname{ind}(H_{\overline{G}}) = 1$ or $D(\overline{G}) = 2$, then $\chi_d(\overline{G}) = D(\overline{G})$.*
2. *If $\operatorname{ind}(H_{\overline{G}}) > 1$ and $D(\overline{G}) > 2$, then $\chi_d(\overline{G}) \leq \operatorname{ind}(H_{\overline{G}})(D(\overline{G})-2)+1$.*

Moreover, the mentioned upper bounds in both cases are sufficient for greedy down-coloring of $\overline{G}$.

REMARKS: (i) Considering the graph $\overline{G}(k,m))$ from Section 2, we have that $H_{\overline{G}(k,m)}$ is a hypergraph with $\operatorname{ind}(H_{\overline{G}(k,m)}) = 2(k-2)$ and $\sigma(H_{\overline{G}(k,m)}) = 2m+1$. By Corollary 1 we have immediately that $\chi_d(\overline{G}(k,m)) \leq 2(k-2)(2m-1)+1 = \Theta(km)$, which agrees with the asymptotic value of the actual down-chromatic number of km, also a $\Theta(km)$ function. Hence, Corollary 1 is asymptotically tight. (ii) Although we have assumed our digraphs to be acyclic, we note that the definition of down-coloring can be easily extended to a regular cyclic digraph $\overline{G}$ by interpreting the notion of descendants of a vertex u to mean the set of nodes reachable from u. In fact, if $\overline{G}$ is an arbitrary digraph, then there is an equivalent acyclic digraph $\overline{G}'$, on the same set of vertices, with an identical down-graph: First form the *condensation* $\hat{G}$ of $\overline{G}$ by shrinking each strongly connected component of $\overline{G}$ to a single vertex. Then form $\overline{G}'$ by replacing each node of $\hat{G}$ which represents a strongly connected component of $\overline{G}$ on a set $X \subseteq V(\overline{G})$ of vertices, with an arbitrary vertex $u \in X$, and then add a directed edge from u to each $v \in X \setminus \{u\}$. This completes the construction. – Observe that each node $v \in X$ has exactly the same neighbors in the down-graph of $\overline{G}'$ as u, as it is a descendant of u and u alone. Further, if node v was in a different strong component of $\overline{G}$ than u but was reachable from u, then it will continue to be a descendant of u in $\overline{G}'$. Hence, the down-graphs of $\overline{G}$ and $\overline{G}'$ are identical.

5 Approximations for down-colorings

In this final section we derive a tight bound of the factor α, as a function of $n = |V(\overline{G})|$, when we can say that the down-coloring of $\overline{G}$ has an α-down-coloring approximation. In what follows, we assume H to be a hypergraph satisfying the unique element property. Unless otherwise stated we therefore have $H = H_{\overline{G}}$.

Theorem 8. *There is a greedy algorithm that approximates the full-coloring of hypergraphs within a factor of $\sqrt{m}$, where m is the number of hyperedges. Thus it approximates the edge-coloring of a hypergraph within a factor of $\sqrt{n}$.*

This yields a $\sqrt{M}$ approximations for the down-coloring of acyclic digraphs, where $M = \max\{\overline{G}\}$ is the number of source vertices in $\overline{G}$.

Proof. Given a hypergraph H, the algorithm finds an inductive order of the vertices $u_1, u_2, \ldots, u_n$, so that for each vertex u_i, the degree of u_i is minimum in the subgraph induced by $U_i = \{u_1, u_2, \ldots, u_i\}$. The algorithm then colors the vertices first-fit in that order. The *inductive degree* of a vertex u_i is its degree in $H[V_i]$, i.e. $\delta(H[V_i])$.

Let $m = |\mathcal{E}(H)|$ and $k = \sqrt{m}$. If $\text{ind}(H) < k$, then the inductive degree of each vertex u_i in $G'[V_i]$ is at most $k(\sigma(H)-1) \le k\chi_s(H)$, and we are done. Thus, assume from now on that $\text{ind}(H) > k$. Let S be the largest prefix set V_i such that the inductive degree of u_i is at least k. Each vertex in $H[S]$ is of degree at least k, thus the average number of vertices per edge is at least $|S|k/|\mathcal{E}(H[S])|$. Hence,

$$\chi_s(H) \ge \sigma(H[S]) \ge |S|k/|\mathcal{E}(H[S])| \ge |S|k/m = |S|/k.$$

The greedy algorithm uses at most one color for each node in S. Thus, if the algorithm does not use more colors, including for vertices in $V - S$, then the performance ratio is at most $|S|/\chi_s(H) \le k$.

Suppose greedy uses more than $|S|$ colors. This implies that some vertex v in $V - S$ had more than $|S|$ neighbors. Since the inductive degree of u in H was less than k it has fewer than $k\sigma(H)$ neighbors in G', and thus u was assigned a color at most $k\sigma(H)$. Since the chromatic number $\chi_s(H)$ is at least $\sigma(H)$, the algorithm maintains a performance ratio of k.

Observe that we bounded the number of colors used by the algorithm in terms of the maximum edge size $\sigma(H)$. Thus, we have shown that the full-chromatic number of a hypergraph H differs from $\sigma(H)$ by a factor of at most $\sqrt{|\mathcal{E}(H)|}$. In terms of down-graphs and hypergraphs, we obtain the following bound.

Corollary 2. $\chi(G') = \chi_s(H) \le \sqrt{|\max\{\overline{G}\}|} \cdot D(\overline{G}) \le \sqrt{n} \cdot D(\overline{G})$.

Compare the above corollary with Observation 2.

We now give a reduction from the ordinary coloring problem that shows that the approximation of the greedy algorithm is close to best possible. Given a graph G_0, we construct an acyclic digraph $\overline{G}$ of height two as follows:

$$\begin{aligned} V(\overline{G}) &= E(G_0) \cup V(G_0), \\ E(\overline{G}) &= \{(e, v) : v \in e, v \in V(G_0), e \in E(G_0)\} \end{aligned}$$

The digraph has a source vertex for each edge in G_0, a leaf vertex for each node in G, and an edge from a source to a leaf if the leaf corresponds to an vertex incident on the edge corresponding to the source vertex.

Let H be the corresponding down-hypergraph. Note that the subhypergraph $H[V(G_0)]$ induced by the leafs is a graph and is exactly the graph G_0. The source nodes of H induce an independent set. Thus, $\chi(G_0) \le \chi_s(H) \le \chi_s(G_0) + 1$. By the results of Feige and Kilian [15], the chromatic number problem cannot be approximated within a factor of $|V(G)|^{1-\epsilon}$, for any $\epsilon > 0$, unless $NP \subseteq ZPP$, i.e. unless there exists polynomial-time randomized algorithms for NP-hard problems. Here, $n = |V(H)| \approx \sqrt{|V(G_0)|}$. Hence, we have the following.

Observation 9 *It is hard to approximate the down-coloring of acyclic digraphs within a factor of $n^{1/2-\epsilon}$, for any $\epsilon > 0$. This holds even for digraphs of height two.*

We may now ask if it is possible to give a better approximation for important special cases of the down-coloring problem. In particular, digraphs arising from family data have some special properties; in particular, each vertex has in-degree at most 2, and normally a fairly small out-degree. We can show that even in this case, we cannot do better.

Theorem 10. *It is hard to approximate the down-coloring of acyclic digraphs within a factor of $n^{1/2-\epsilon}$, for any $\epsilon > 0$, even when restricted to acyclic digraphs of in-degree and out-degree two.*

Proof. Recall that the digraph $\overline{G}$ has maximum out-degree two. We modify it into an acyclic digraph $\overline{G}_1$ that has also maximum in-degree two. We replace each sink u in $\overline{G}$ by a full binary tree T_u directed towards u. There is a sink w_v in T_u for each node v such that $(v, u) \in E(\overline{G})$, with an arc (v, w_v) added to $\overline{G}_1$. This completes the construction. Observe that $\overline{G}_1$ has the same set of sources and sinks as $\overline{G}$, and additionally $\sum_{v \in G} 2d(v) - 2 = 4|E(G)| - |V(G)|$ internal nodes.

The height of a node v in $\overline{G}_1$ is the distance from u to a sink; the maximum height is $\lg \Delta(G) \leq \lg n$. Let x and y be nodes of the same height in $\overline{G}$ and let v_x and v_y denote the sinks in whose trees x and y belong. We observe that if x and y are adjacent in the down-graph of $\overline{G}_1$, then v_x and v_y are distinct and adjacent in the down-graph. Namely, incomparable nodes in the same tree do not have common ancestors; also, since v_x and v_y are descendants of x and y, they will be adjacent if x and y are. This implies, that the chromatic number of the subgraph induced by non-source nodes of the same height is at most the chromatic number of the subgraph induced by the sinks, or $\chi(G)$. The source nodes of $\overline{G}_1$ still form an independent sets, while the other nodes are of $\lg n$ heights. Thus, the chromatic number of the down-graph of $\overline{G}_1$ is at most $(\lg n + 1)\chi(G)$, but also at least $\chi(G)$. Since it is hard to approximate the chromatic number of G within $|V(G)|^{1-\epsilon}$ factor, for any $\epsilon > 0$, and $n = |V(\overline{G}_1)| = O(|V(G)|^2)$, it is also hard to approximate the chromatic number of $\overline{G}_1$ within $n^{1/2-\epsilon}$, for any $\epsilon > 0$.

Acknowledgments

The authors are grateful to John L. Pfaltz for his interest and encouragements to submit this article.

References

1. Geir Agnarsson and Ágúst Egilsson. *On vertex coloring simple genetic digraphs.* submitted.
2. The Gene Ontology Consortium. *Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000).* Nature Genet, **25**, 25 – 29, (2000).

3. Serge Abitebboul, Peter Buneman and Dan Suciu. *Data on the Web, From Relations to Semistructured Data and XML.* Morgan Kaufmann Publishers (2000).
4. An Oracle White Paper. *Key Data Warehousing Features in Oracle9i: A Comparative Performance Analysis.* Available online from Oracle at `http://otn.oracle.com/products/oracle9i/pdf/o9i_dwfc.pdf`, (September 2001).
5. Michael W. Cain (iSeries Teraplex Integration Center), *Star Schema Join Support within DB2 UDB for iSeries Version 2.1.* Available online from IBM at `http://www-919.ibm.com/developer/db2/documents/star/`, (October 2002).
6. B. Bhattacharjee, L. Cranston, T. Malkemus, and S. Padmanabhan. *Boosting Query Performance: Multidimensional Clustering* DB2 Magazine, Quarter 2, 2003, Vol. 8, Issue 2. Also, available online at `http://www.db2mag.com`
7. Clifford A. Shaffer. *A Practical Introduction to Data Structures and Algorithm Analysis* Java Edition, Prentice Hall, (1998).
8. C. C. Harner and R. C. Entringer. *Arc colorings of digraphs.* Journal of Combinatorial Theory, Series B, **13**, 219 – 225, (1972).
9. H. Jacob and H. Meyniel. *Extensions of Turán's Brooks' theorems and new notions of stability and colorings in digraphs.* Combinatorial Mathematics, North-Holland Math. Stud., **75**, 365 – 370, (1983).
10. Xiang Ying Su. *Brooks' theorem on colorings of digraphs.* Fujian Shifan Daxue Xuebao Ziran Kexue Ban, **3**, no. 1, 1 – 2, (1987).
11. Douglas B. West. *Introduction to Graph Theory.* Prentice Hall, second edition, (2001).
12. William T. Trotter. *Combinatorics and Partially Ordered Sets, Dimension Theory.* Johns Hopkins Series in the Mathematical Sciences, The Johns Hopkins University Press, (1992).
13. Weifan Wang and Kemin Zhang. *Colorings of hypergraphs.* Adv. Math. (China), **29**, no. 2, 115–136, (2000).
14. Geir Agnarsson and Magnús M. Halldórsson, *Coloring Powers of Planar Graphs*, SIAM Journal of Discrete Mathematics, to appear.
15. U. Feige and J. Kilian, *Zero Knowledge and the Chromatic number.* Journal of Computer and System Sciences, **57**, no 2, 187–199, (1998).

May 30, 2003