Colouring random intersection graphs and complex networks

Michael Behrisch [*]

Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

Anusch Taraz [†]

Zentrum Mathematik, Technische Universität München, 80290 München, Germany

Michael Ueckerdt

Institut für Informatik, Humboldt-Universität zu Berlin, 10099 Berlin, Germany

December 8, 2005

Random intersection graphs naturally exhibit a certain amount of transitivity and hence can be used to model real–world networks. We study the evolution of the chromatic number of a random intersection graph and show that, in a certain range of parameters, these random graphs can be coloured optimally with high probability using different greedy algorithms. Experiments on real network data confirm the positive theoretical predictions and suggest that heuristics for the clique and the chromatic number can work hand in hand proving mutual optimality.

1 Introduction and results

The classical random graph model, introduced by Erdős and Rényi in the early 1960s, considers a fixed set of n vertices and edges that exist with a certain probability $p = p(n)$, independently from each other. It was shown to be inappropriate for describing real–world networks because it lacks certain features of those such as a scale free degree distribution and the emergence of local clusters. One of the underlying reasons that are responsible for this mismatch is precisely the independence of the edges, in other words the missing transitivity: if vertices x and y exhibit a relationship of some kind in a real–world network and so do vertices y and z, then this suggests a connection between vertices x and z, too.

Intersection graphs. Suppose that we have a vertex set V and another set W. An *intersection graph* is a graph with vertex set V, where we assign to each vertex

[*]supported by the DFG research center MATHEON in Berlin.

[†]supported in part by the DFG research center MATHEON in Berlin.

v a subset $W_v \subseteq W$ and connect two vertices v, v' by an edge if and only if their assigned sets W_v and $W_{v'}$ have non-empty intersection.

We call the ground set W from which the assigned sets are chosen *universal feature set* and its elements *features*. If feature $w \in W_v$, we say that feature w is *assigned* to vertex v or simply that v *has* w. The set W_v is called *feature set* of v. For a specified $w \in W$, let V_w be the set of vertices v that have feature w. We call V_w a *feature clique*, since it obviously induces a clique in the intersection graph. As usual, $\Gamma(v)$ denotes the set of neighbours of v, i.e. the set of vertices in V that have features with v in common.

Well studied examples for intersection graphs are interval graphs on the real line. In this paper, however, we will only consider finite sets. Obviously every graph is an intersection graph (simply pick an individual feature assigned only to the two vertices of every edge), but the fewer features we have, the more apparent becomes the structure of the shared features inside the graph.

Random intersection graphs. A *random intersection graph* on n vertices with a universal feature set W of size m is a random graph with vertex set $[n]$ where each vertex gets assigned a random set of features by choosing each feature independently with probability p. A sample of this probability space is denoted by $G_{n,m,p}$. We consider now and in the following $m := n^\alpha$, and will usually distinguish two cases: $\alpha > 1$ and $0 < \alpha < 1$. If the probability of $G_{n,m,p}$ having a property $\mathcal{A}$ tends to 1 with n tending to infinity, we say that $G_{n,m,p}$ has property $\mathcal{A}$ *asymptotically almost surely* (a.a.s.).

It is sometimes convenient to look at the random intersection graph as a random bipartite graph with bipartition (V, W) and edges occurring between the two classes independently with probability p. Such a graph will be called a *generator*.

Several aspects of random intersection graphs have been studied before. Karoński, Scheinerman, and Singer-Cohen [11] study subgraph appearance in this model. Fill, Scheinerman, and Singer-Cohen [5] investigate the equivalence of $G_{n,m,p}$ to $G_{n,p}$, and Stark [14] analyses its vertex degree distribution. Behrisch and Taraz [3] show how to reconstruct the feature structure when only the random intersection graph is given as input. A study of the component evolution is given by Behrisch in [2]. Some results concerning connectivity and cliques can be found in Singer [13]. Extensions to the model are proposed by Godehardt and Jaworski in [7], who modify the distribution of the sizes of the feature cliques. The practical relevance of random intersection graphs is studied by Newman, Strogatz and Watts in [12] and by Guillaume and Latapy in [9].

The aim of this paper is to investigate the evolution of the chromatic number of $G_{n,m,p}$. As usual, denote by $\chi(G)$ the chromatic number of G and by $\omega(G)$ the size of the largest clique in G. The computation of these two fundamental parameters is long known to be NP-hard. Our main results are that for a random intersection graph $G = G_{n,m,p}$ where m and p lie in a certain range, asymptotically almost surely $\chi(G)$ and $\omega(G)$ can be computed efficiently by simple colouring heuristics and actually coincide.

Theorem 1. *Let $m := n^\alpha$ with $\alpha > 0$ fixed and $p \ll \sqrt{\frac{1}{nm}}$. Then $G_{n,m,p}$ can a.a.s. be coloured optimally in linear time and $\chi(G_{n,m,p}) = \omega(G_{n,m,p})$.*

Theorem 2. *Let $m := n^\alpha$ with $0 < \alpha < 1$ fixed and $p \ll \frac{1}{m \ln n}$. Then $G_{n,m,p}$ can a.a.s. be coloured optimally in linear time. Moreover, for $np > \ln^4 n$ we have a.a.s.*

$$\chi(G_{n,m,p}) = \omega(G_{n,m,p}) \ \sim np.$$

Note that in principle one could also state in Theorem 1 that for $np > \ln^4 n$ we have a.a.s. $\chi(G_{n,m,p}) = \omega(G_{n,m,p}) \ \sim np$, but this is redundant since $np > \ln^4 n$ and $p \ll \sqrt{\frac{1}{nm}}$ together imply $\alpha < 1$ and thus the two theorems overlap in this case.

Applications. We started this section by claiming that the $G_{n,m,p}$-model provides a good approximation of real-world networks. Indeed, we have tested our colouring heuristics on real–world networks from application areas such as the internet, co-operation graphs and protein databases. In many cases, the experimental and the theoretical results agree with each other – see Section 4 for details, in particular Figure 1 for a graphical representation describing the parameters in the theorems and how they relate to the experimental results. Still the question remains, *why* one should try to *colour* complex networks. Of course, knowledge of the chromatic number gives important structural information of a general nature, but while for instance the clique number is practically meaningful – the size of the largest cluster in the network – the chromatic number seems to be of less immediate use.[1]

There is however one important application of the chromatic number, and this is exactly the clique number. Suppose we have a heuristic that tries to find the maximal size of a clique. If we also have a heuristic that tries to determine the minimum number of colours, and both of the proposed numbers coincide (or are at least very close to each other), then this proves that both numbers have already reached (near-) optimal values. This is precisely what we did in our experiments: we applied different heuristics discussed in an earlier paper [3] to find large cliques (and good clique covers) in the networks. At the same time, we tried to find good colourings of real–world networks using the greedy algorithms discussed in this paper. The results showed that, just as predicted by Theorems 1 and 2, the proposed chromatic number and clique number indeed coincide (or are at least very close to each other).

In a way this is very reminiscent of the theory of perfect graphs. In fact, $G_{n,m,p}$ with m and p as in Theorem 1 is a.a.s. perfect, and we can thus use some of the perfect graph methodology to give a short proof of the theorem. For parameters m and p as in Theorem 2, although $\chi(G_{n,m,p}) = \omega(G_{n,m,p})$ a.a.s., $G_{n,m,p}$ is not perfect and hence a different colouring strategy has to be used for this case.

The paper is organised as follows. After a short section containing some auxiliary tools, we will prove Theorems 1 and 2 in Sections 3.1 and 3.2 respectively. Our colouring experiments can be found in Section 4, and a brief outlook concludes the paper.

[1] One possible application, not to be taken too seriously, could be to distribute film-stars to a minimum number of hotels (colour classes) in such a way that co-stars of the same movie are not put in the same hotel, just to avoid trouble.

2 Auxiliary Lemmas

The following estimates are used without proof:

$$1 - ab \le (1-a)^b \le 1 - \frac{ab}{2} \qquad \text{for } 0 \le a \le 1, ab < 1. \tag{1}$$

Let X be a non-negative random variable with expectation $\mu = \mathbb{E}[X]$. As a special case of Markov's inequality the first moment method states that

$$\mathbb{P}[X \ge 1] \le \mu. \tag{2}$$

If X is binomially distributed random variable (n trials, each with probability p), then $\mu = np$ and we shall use the following variants of Chernoff's inequality (see Section 2 in [10]):

$$\mathbb{P}[X \ge \mu + t] \le \exp\left(-\frac{t^2}{2(\mu + t/3)}\right) \qquad \text{for } t \ge 0, \tag{3}$$

$$\mathbb{P}[X \le \mu - t] \le \exp\left(-\frac{t^2}{2\mu}\right) \qquad \text{for } t \ge 0, \tag{4}$$

$$\mathbb{P}[X \ge t] \le \exp(-t) \qquad \text{for } t \ge 7\mu. \tag{5}$$

We first show that the probability that there is a feature clique in $G_{n,m,p}$ which deviates much from its expected size is exponentially small.

Lemma 3. *Let $X_w := |V_w|$ be the random variable counting the number of vertices of a fixed feature w in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then*

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}}\right] \le m \exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right).$$

Proof. The number of vertices chosen by a feature is a binomially distributed variable. Its deviation from its expected value can therefore be bounded by Chernoff inequalities (3) and (4). First let w be fixed:

$$\mathbb{P}\left[X_w > pn + (pn)^{\frac{3}{4}}\right] \le \exp\left(-\frac{(pn)^{\frac{3}{2}}}{2(pn + (pn)^{\frac{3}{4}}/3)}\right) \le \frac{1}{2}\exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right)$$

$$\mathbb{P}\left[X_w < pn - (pn)^{\frac{3}{4}}\right] \le \exp\left(-\frac{(pn)^{\frac{3}{2}}}{2pn}\right) \le \frac{1}{2}\exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right).$$

By linearity of expectation (summing over all possible w) and Markov's inequality this implies that

$$\mathbb{P}\left[\exists w \in W : |X_w - pn| > (pn)^{\frac{3}{4}}\right] \le m \exp\left(-\frac{(pn)^{\frac{1}{2}}}{3}\right).$$

□

Since we are mostly interested in small feature sets, we need only an upper bound on their size.

Lemma 4. *Let $X_v := |W_v|$ be the random variable counting the number of features for a fixed vertex v in a random intersection graph $G_{n,m,p}$ with $m := n^\alpha$ and $\alpha < 1$. Then for $pm \leq 3 \ln n$*

$$\mathbb{P}\left[\exists v \in V : X_v > 21 \ln n\right] \leq \frac{1}{n^{20}}.$$

Proof. Very similarly to the previous lemma, we have for a fixed vertex v and for $pm \leq 3 \ln n$

$$\mathbb{P}\left[X_v > 21 \ln n\right] \overset{(5)}{\leq} \exp(-21 \ln n) = \frac{1}{n^{21}}.$$

Again summing over all vertices v yields the statement of the lemma. □

3 Proofs

In the following two subsections we describe two simple and well known deterministic algorithms that find a proper colouring of a given input graph $G = (V, E)$ in linear time. Both algorithms are greedy heuristics: they colour the vertices in a prescribed order and assign to each vertex the smallest colour that has not been used for any of its neighbours which are already coloured. Thus the main task is to prove the following: if the input graph G is a random intersection graph $G_{n,m,p}$ with parameters n, m and p as given in Theorems 1 and 2, then these algorithms will asymptotically almost surely produce a colouring with (at most) $\omega(G)$ different colours. Hence the colouring is optimal and $\chi(G) = \omega(G)$, as required.

The additional claim in Theorem 2 that a.a.s. $\omega(G)$ is of order np will follow from the fact that the largest clique is a feature clique, which according to Lemma 3 is of that order.

3.1 Perfect elimination scheme

The aim of this subsection is to prove Theorem 1. Here is the basic idea of our colouring algorithm. We first try to order the vertices of the graph as $x_n, \ldots, x_1$ in such a way that for every vertex x_i the 'remaining neighbourhood' $\Gamma(x_i) \cap \{x_{i-1}, \ldots, x_1\}$ induces a clique in G. Having established this ordering, we greedily colour the vertices in the (reverse) order $x_1, \ldots, x_n$. Observe that this implies that vertices which are contained in many different cliques, e.g. those that have many features, will be coloured relatively early.

Such an ordering is called a *perfect elimination scheme*, in short *PES*. Tarjan and Yannakakis [15] proved that, if a graph has a PES, a so-called maximum cardinality search will produce a PES in linear time. If the graph doesn't have a PES, then the procedure returns an arbitrary ordering. This leads to the following greedy colouring heuristic:

Algorithm 1.
Input: *Graph $G = (V, E)$ on n vertices*

Output: *colouring of G*
GreedyColourPES*(G)*
(1) $A := \emptyset$
(2) ***for*** $i := 1$ ***to*** n
(3) *choose* $x_i \in V \setminus A$ *such that* $|\Gamma(x_i) \cap A|$ *is maximal*
(4) $A := A + x_i$
(5) ***for*** $i := 1$ ***to*** n
(6) *colour* x_i *with the smallest colour not occurring in* $\Gamma(x_i)$

The following three crucial facts have been known for a long time:

1. a graph G has a PES (and it can be found in linear time and can be found as described above) if and only if G is *chordal*, i.e. it does not contain an induced cycle with more than three vertices [15],
2. chordal graphs are perfect [4, Chapter 5.5], thus in particular $\chi(G) = \omega(G)$, and
3. if a PES exists for G, then using it as described above the greedy colouring procedure colours G optimally.

The last observation is a folklore result and obviously true: if the set of the already coloured neighbours of every vertex x_i forms a clique when x_i is coloured, then whenever a vertex x_i needs a new colour k, we have just found a clique of size k, and hence k colours are really needed to colour the graph.

Now all that remains to do is to prove that $G_{n,m,p}$ is chordal for the given parameters n, m and p, which will be done in the following lemma.

Lemma 5. *Let* $m := n^{\alpha}$ *for* $\alpha > 0$ *fixed and* $p \ll \sqrt{\frac{1}{nm}}$. *Then* $G_{n,m,p}$ *is a.a.s. chordal.*

Proof. Let $G = G_{n,m,p}$ be a random intersection graph and $B = (V \cup W, E_B)$ a bipartite generator of G. By definition, G is chordal iff it does not contain an induced cycle of length at least four. Suppose that $v_1, \ldots, v_k$ form an induced cycle C_k in G. Then there must exist features $w_1, \ldots, w_k$ such that w_i is a feature of both v_i and v_{i+1} for all $i \in [k-1]$, and w_k is a feature for both v_k and v_1. Moreover all the w_i are distinct, since otherwise the cycle wouldn't be induced. This yields a cycle $v_1, w_1, v_2, w_2, \ldots, v_k, w_k$ in the generator B. The probability for such a cycle in B can obviously be bounded from above by p^{2k}, and multiplying this with the number of possibilities to choose $v_1, \ldots, v_k$ and $w_1, \ldots, w_k$ we get:

$$\mathbb{P}\left[G \text{ contains an induced } C_k\right] \;\le\; n^k m^k p^{2k} \;=\; (nmp^2)^k.$$

The probability of G being not chordal is now bounded by:

$$\begin{aligned}
\mathbb{P}\left[G \text{ is not chordal}\right] &\leq \sum_{k=4}^{min(n,m)} \mathbb{P}\left[G \text{ contains an induced } C_k\right] \\
&\leq \sum_{k=4}^{min(n,m)} (nmp^2)^k \\
&\leq \sum_{k=0}^{\infty} (nmp^2)^k - 1 \;=\; \frac{1}{1-nmp^2} - 1,
\end{aligned}$$

which tends to 0 for n tending to infinity because nmp^2 tends to 0. $\square$

A second moment calculation (see Singer [13]) shows that $p = \sqrt{\frac{1}{nm}}$ is in fact the threshold function for the appearance of induced cycles of *fixed* length $k \geq 4$ in random intersection graphs. Thus for $p \gg \sqrt{\frac{1}{nm}}$ these graphs are a.a.s. not chordal.

3.2 Smallest last heuristic

The aim of this subsection is to prove Theorem 2. Again we employ a greedy strategy but this time the precomputed ordering $x_1, \ldots, x_n$ of the vertices is slightly different. Suppose we have already selected $x_n, \ldots, x_{i+1}$. Then among the remaining vertices x_i is the vertex with the smallest number of neighbours (among the remaining vertices). More precisely:

Algorithm 2.
Input: *Graph $G = (V, E)$ on n vertices*
Output: *colouring of G*
GreedyColourSmallestLast*(G)*
(1) $A := V$
(2) ***for*** $i := n$ ***downto*** 1
(3) *choose $x_i \in A$ such that $|\Gamma(x_i) \cap A|$ is minimal*
(4) $A := A - x_i$
(5) ***for*** $i := 1$ ***to*** n
(6) *colour x_i with the smallest colour not occurring in $\Gamma(x_i)$*

As there may be more than one such ordering, we denote by $\chi_{\mathrm{SL}}(G)$ the maximum number of colours that GreedyColourSmallestLast(G) uses for an input graph G. It is well known [4, Chapter 5.2] that the number of colours used by the algorithm is always bounded from above by the maximal minimum degree of all subgraphs of G, plus one:

$$\chi_{\mathrm{SL}}(G) \;\leq\; 1 + \max_{H \subseteq G} \delta(H). \tag{6}$$

From this we derive the following simple proposition.

Proposition 6. *If G is a graph such that*

$$\text{every vertex } v \text{ has less than } \omega(G) \text{ neighbours of degree at least } \omega(G), \tag{7}$$

then

$$\chi_{SL}(G) = \omega(G) = \chi(G).$$

Proof. We claim that (7) implies that

$$1 + \max_{H \subseteq G} \delta(H) \leq \omega(G). \tag{8}$$

Suppose for a contradiction that there exists a subgraph H with $1+\delta(H) > \omega(G)$. Let v be a vertex of minimal degree in H, i.e. $d_H(v) = \delta(H) \geq \omega(G)$. Then for *all* neighbours w of v in H we have

$$d_G(w) \geq d_H(w) \geq d_H(v) = \delta(H) \geq \omega(G),$$

and since there are $d_G(v) \geq d_H(v) = \delta(H) \geq \omega(G)$ neighbours of v in G, this contradicts the property in (7), which proves the claim in (8).

Now we are done, since

$$\chi(G) \leq \chi_{\mathrm{SL}}(G) \overset{(6)}{\leq} 1 + \max_{H \subseteq G} \delta(H) \overset{(8)}{\leq} \omega(G) \leq \chi(G).$$

$\square$

Let us move back to intersection graphs. In the following we call a vertex v *rich* if it has at least two features. Obviously, the only way that a vertex can have degree at least $\omega(G)$ is if it is rich. Hence we have the following corollary.

Corollary 7. *Suppose that G is an intersection graph such that every vertex has less than $\omega(G)$ rich neighbours, then*

$$\chi_{SL}(G) = \omega(G) = \chi(G).$$

$\square$

In order to prove that in our random intersection graph, the condition of the above corollary is a.a.s. satisfied, we first obtain an upper bound on the number of rich vertices in each feature clique.

Lemma 8. *Let $m = n^\alpha$ for $0 < \alpha < 1$ fixed, $p \geq \frac{10 \ln^2 n}{n}$ and $t \geq 0$. Denote by ω_f the size of a largest feature clique in $G_{n,m,p}$. Then in a random intersection graph $G_{n,m,p}$ the probability that there exists a feature clique C with more than $\omega_f mp + t$ rich vertices is at most*

$$m \ \exp\left(-\frac{t^2}{2\omega_f mp + 2t/3}\right)$$

Proof. Let $C \subseteq V$ denote an arbitrary feature clique in G. For $v \in C$ we denote by $X_{C,v}$ the random variable which is 1 whenever v is rich and 0 otherwise. Then

$$\mathbb{P}\left[X_{C,v} = 1\right] = 1 - (1-p)^{m-1} \overset{(1)}{\leq} 1 - (1 - (m-1)p) \leq mp.$$

Let $X_C := \sum_{v\in C} X_{C,v}$ count the rich vertices in C. For the expectation of X_C we have:

$$\mathbb{E}[X_C] \;=\; \sum_{v\in C} \mathbb{P}[X_{C,v}=1] \;\leq\; \omega_f mp.$$

Using the Chernoff bound we get:

$$\begin{aligned}\mathbb{P}[X_C \geq \omega_f mp + t] \;&\leq\; \mathbb{P}[X_C \geq \mathbb{E}[X_C] + t]\\ &\overset{(3)}{\leq}\; \exp\left(-\frac{t^2}{2\mathbb{E}[X_C]+2t/3}\right) \;\leq\; \exp\left(-\frac{t^2}{2\omega_f mp + 2t/3}\right).\end{aligned}$$

Of course the events '$X_C \geq \omega_f mp + t$' are not independent of each other for overlapping feature cliques C, but using linearity of expectation and the Markov inequality (2) we can bound the probability of existence of a feature clique with too many rich vertices by the expression in the lemma. □

Proof of Theorem 2. We want to apply Corollary 7 and hence need to show that in $G = G_{n,m,p}$ every vertex has less than $\omega(G)$ rich neighbours. Recall that $m := n^\alpha$ with $0 < \alpha < 1$ fixed and $p \ll \frac{1}{m\ln n}$. First observe that we can assume that $pn > \ln^4 n$, since otherwise p would be so small that we could apply Theorem 1 instead. Set

$$t := \max(3\ln n, \sqrt{nmp^2}\ln n),$$

and consider an arbitrary small $\varepsilon > 0$. We shall make use of the following two technical observations (involving t) that will be verified later:

$$21\ln n((1+\varepsilon)nmp^2 + t) \leq (1-\varepsilon)np, \tag{9}$$

$$m\exp\left(-\frac{t^2}{2(1+\epsilon)nmp^2 + 2t/3}\right) \leq n^{\alpha-1}. \tag{10}$$

Again denote by ω_f the size of a largest feature clique in $G = G_{n,m,p}$ and consider the following events that have already been discussed in Lemmas 3, 4 and 8 respectively:

$\mathcal{A}$: for all $w \in W : ||V_w| - pn| < \varepsilon pn$,

$\mathcal{B}$: for all $v \in V : |W_v| \leq 21\ln n$,

$\mathcal{C}$: every feature clique C has at most $\omega_f mp + t$ rich vertices.

Let Y_v be the number of rich neighbours of a vertex v. Then Y_v is bounded from above by the number of feature cliques containing v, multiplied with the number of rich vertices per feature clique, and we can then compare this to the size of a feature clique, which is a lower bound for $\omega(G)$. So if all the events $\mathcal{A},\mathcal{B},\mathcal{C}$ hold, then

$$Y_v \leq 21\ln n\ ((1+\varepsilon)pn\ mp + t) \overset{(9)}{\leq} (1-\varepsilon)np \overset{(\mathcal{A})}{<} \omega_f - 1 < \omega(G), \tag{11}$$

which would immediately prove (most of) the statements in Theorem 2 because of Corollary 7. To prove that $\omega(G) \sim np$, note that by the estimate in (11) there is no vertex v with $\omega_f - 1$ rich neighbours, and hence there exists no clique of size ω_f containing only rich vertices. In turn, this implies that $\omega(G) = \omega_f$, since a clique which is not (subset of) a feature clique contains only rich vertices, and we are done because $\omega_f \sim np$ by property $\mathcal{A}$.

Let us complete the proof by showing that a.a.s. all the events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ hold. Obviously

$$\mathbb{P}\left[\mathcal{A}\cap\mathcal{B}\cap\mathcal{C}\right] = 1-\mathbb{P}\left[\bar{\mathcal{A}}\right]-\mathbb{P}\left[\mathcal{A}\cap\bar{\mathcal{B}}\right]-\mathbb{P}\left[\mathcal{A}\cap\mathcal{B}\cap\bar{\mathcal{C}}\right] \geq 1-\mathbb{P}\left[\bar{\mathcal{A}}\right]-\mathbb{P}\left[\bar{\mathcal{B}}\right]-\mathbb{P}\left[\mathcal{A}\cap\bar{\mathcal{C}}\right],$$

so it suffices to check that all the probabilities $\mathbb{P}\left[\bar{\mathcal{A}}\right], \mathbb{P}\left[\bar{\mathcal{B}}\right], \mathbb{P}\left[\mathcal{A}\cap\bar{\mathcal{C}}\right]$ tend to zero. For the first two this is immediately implied by Lemma 3 (which applies because of $m < n$ and $pn > \ln^4 n$) and Lemma 4 respectively. For the latter it follows from Lemma 8 and observing that

$$\mathbb{P}\left[\bar{\mathcal{A}}\cap\mathcal{C}\right] \leq m\ \exp\left(-\frac{t^2}{2(1+\varepsilon)pn\ mp+2t/3}\right) \overset{(10)}{\leq} n^{\alpha-1},$$

which does tend to zero, since $\alpha < 1$.

Thus all that remains to be done is to check the two technical observations (9) and (10). Considering (9), we distinguish two cases. For $\sqrt{nmp^2} > 3$ we have

$$\begin{aligned} 21\ln n((1+\varepsilon)nmp^2 + \sqrt{nmp^2}\ln n) &\leq 40nmp^2\ln n + 21\sqrt{nmp^2}\ln^2 n \\ &= np(40mp\ln n + 21\sqrt{m/n}\ln^2 n). \end{aligned}$$

which is smaller than $(1-\varepsilon)np$ because of $mp \ll \frac{1}{\ln n}$ and $\alpha < 1$.

And for $\sqrt{nmp^2} \leq 3$

$$\begin{aligned} 21\ln n((1+\varepsilon)nmp^2 + 3\ln n) &\leq 40nmp^2\ln n + 63\ln^2 n \\ &\leq 360\ln^3 n + 63\ln^2 n. \end{aligned}$$

which is smaller than $(1-\varepsilon)np$ because of $\frac{\ln^3 n}{n} \ll p$.

Considering (10), we distinguish two cases again. For $\sqrt{nmp^2} > 3$ we have

$$\begin{aligned} m\exp\left(-\frac{nmp^2\ln^2 n}{2(1+\epsilon)nmp^2 + \frac{2}{3}\sqrt{nmp^2}\ln n}\right) &\leq m\exp\left(-\frac{nmp^2\ln^2 n}{nmp^2\ln n}\right) \\ &= m\exp\left(-\ln n\right) = n^{\alpha-1}. \end{aligned}$$

and for $\sqrt{nmp^2} \leq 3$

$$\begin{aligned} m\exp\left(-\frac{9\ln^2 n}{2(1+\epsilon)nmp^2 + \frac{2}{3}3\ln n}\right) &\leq m\exp\left(-\frac{9\ln^2 n}{100 + 2\ln n}\right) \\ &\leq m\exp\left(-\ln n\right) = n^{\alpha-1}. \end{aligned}$$

$\square$

4 Experiments

We have tested our algorithms on eight real–world networks from different application areas. The first five graphs are the same as in [9]. “Internet” describes part

	Internet ▼	Web ■	Authors ⧫	Actors ⋆	Proteins ♠
n	75885	325729	16400	392340	2113
$\|E\|$	357317	1090108	29552	15038083	2203
α	1.1049	1.0210	0.9653	0.9129	0.9886
$\log_n p$	-0.9527	-0.9356	-0.9166	-0.7880	-0.9463
Greedyχ	22	155	11	294	6
GreedyPESχ	21	156	8	294	6
GreedySLχ	20	155	8	294	6
largest clique	20	155	8	294	6

	Mercator ▲	DIP ♣	Drugs •
n	284805	5119	2000
$\|E\|$	449246	14434	163969
α	1.0200	0.9488	0.3713
$\log_n p$	-0.9643	-0.8731	-0.3197
Greedyχ	38	42	381
GreedyPESχ	33	42	381
GreedySLχ	33	42	381
largest clique	19	42	381

Table 1: Statistics on the performance of the algorithms on eight real–world networks

of the internet computer network, “Web” is the link graph of a complex website, “Authors” denotes a coauthoring graph, “Actors” denotes a costarship graph of actors as found in the internet movie database, and “Proteins” is an interaction graph of proteins. For details see [9] and [1]. The “Mercator” graph is a graph of the internet at router level taken from [8]. Moreover “DIP” stands for “Dictionary of Interfaces in Proteins” and is a similarity graph of protein parts (vertices are protein interfaces that are adjacent if they are similar) studied in [6]. “Drugs” is the result of a search for “relatives” of 13 substances in a database of 2000 drugs where an edge connects a pair of drugs which are relatives to the same test substance. Details concerning this network are described in [16].

These are the graphs for which we tried to find good colourings. Greedyχ, GreedyPESχ and GreedySLχ denote the number of colours needed by a greedy colouring procedure that colours the vertices in the natural order (in which they were read), in a PES ordering (cf Algorithm 1) and in a smallest last ordering (cf Algorithm 2) respectively. The following table also states the size of the largest clique we were able to find in the graphs using the clique cover algorithm described in [3]. Obviously the difference between the proposed number of colours and the proposed size of a largest clique is an upper bound of the distance of either number to the optimal value.

The results show that the smallest last heuristic seems to perform well on real–world graphs. In seven cases we were able to colour the graph optimally using the heuristic described. Figure 1 gives a graphical representation of the parameter ranges of Theorems 1 and 2 and shows that, as illustrated by the positions of our example networks, the algorithms work well even outside these ranges.

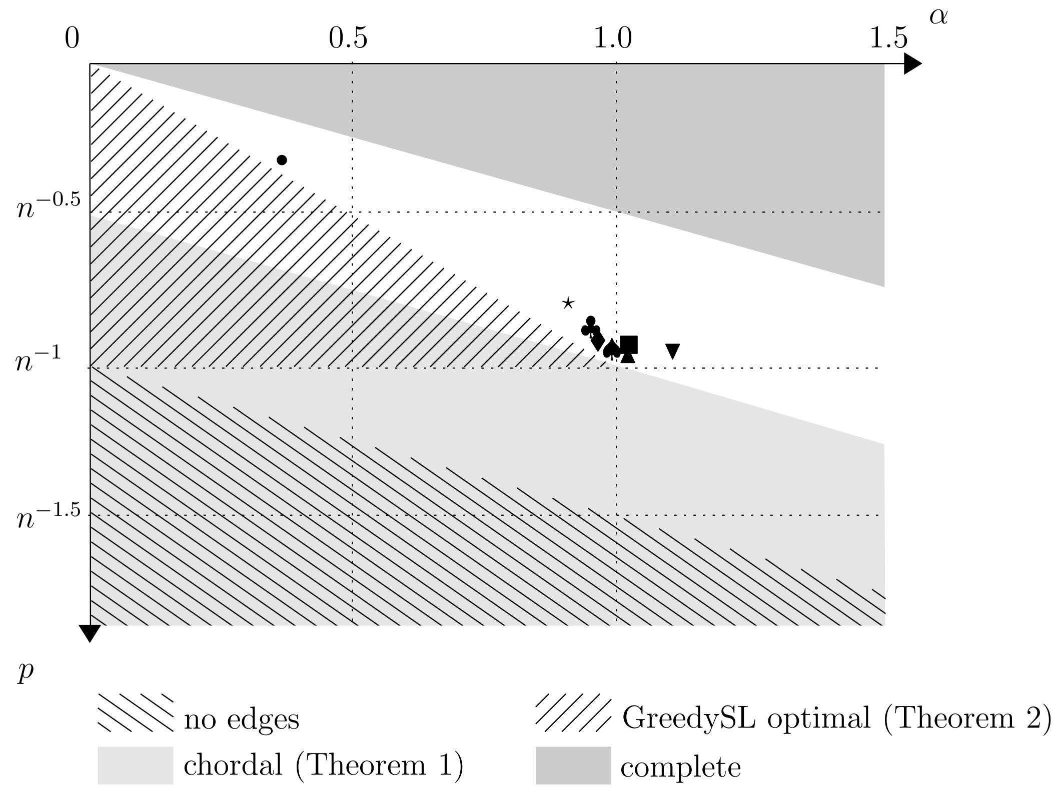

Figure 1: Ranges for p and α where we colour optimally and experimental results

5 Outlook

For the ranges not covered by Theorems 1 and 2, the chromatic number seems to be more difficult to estimate. From the aforementioned result by Singer [13] it is clear that those graph are no longer chordal for $p \gg \sqrt{\frac{1}{nm}}$ while the results on the clique cover [3] suggest that the feature cliques stay the dominant structural element up to $p < \min\{\frac{1}{5}m^{-\frac{2}{3}}, \frac{n}{8m^2}\}$.

In higher ranges, the approximation of the chromatic number by the size of the largest feature clique will not be very good. Using a different approach [17], we tried to establish a better lower bound via the independence number. Using the fact that the chromatic number of any graph is at least as high as the number of vertices divided by the size of a largest independent set, we obtain a lower bound on the chromatic number which beats the size of the largest feature clique, as the following result shows.

Theorem 9 ([17])**.** *Let $\varepsilon > 0$ be fixed and let $m := n^{\alpha}$ with $\alpha > 0$ fixed and $\frac{\ln n}{m} \ll p \ll \sqrt{\frac{\ln n}{m}}$. Then a.a.s. the random intersection graph $G_{n,m,p}$ has no independent set of size*

$$(2 + \varepsilon)\frac{\ln n}{mp^2},$$

which implies that

$$\chi(G_{n,m,p}) \geq \frac{p^2mn}{(2+\varepsilon)\ln n} \gg pn.$$

Lower bounds on the independence number (which match the upper bounds by a logarithmic factor) can also be found in [17].

References

[1] R. Albert, H. Jeong, and A.-L. Barabási. Database of self-organized networks. http://www.nd.edu/ networks/database/index.html.

[2] M. Behrisch. Component evolution in random intersection graphs. Preprint, November 2004.

[3] M. Behrisch and A. Taraz. Efficiently covering complex networks with cliques of similar vertices. *Theoretical Computer Science*, 2005. to appear.

[4] R. Diestel. *Graph theory.* Springer, New York, 1997.

[5] J. A. Fill, E. R. Scheinerman, and K. B. Singer-Cohen. Random intersection graphs when $m = \omega(n)$: An equivalence theorem relating the evolution of the $G(n, m, p)$ and $G(n, p)$ models. *Random Structures and Algorithms*, 16(2):156–176, March 2000.

[6] C. Frömmel, C. Gille, A. Goede, C. Gröpl, S. Hougardy, T. Nierhoff, R. Preissner, and M. Thimm. Accelerating screening of 3D protein data with a graph theoretical approach. *Bioinformatics*, 19(18):2442–2447, 2003.

[7] E. Godehardt and J. Jaworski. Two models of random intersection graphs and their applications. *Electronic Notes in Discrete Mathematics*, 10, 2001.

[8] R. Govindan and H. Tangmunarunkit. SCAN+Lucent internet map from the ISI, November 1999. http://www.isi.edu/div7/scan/mercator/maps.html.

[9] J.-L. Guillaume and M. Latapy. Bipartite structure of all complex networks. *Information Processing Letters*, 90:215–221, 2004.

[10] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs.* John Wiley & Sons, 2000.

[11] M. Karoński, E. R. Scheinerman, and K. B. Singer-Cohen. On random intersection graphs: The subgraph problem. *Combinatorics, Probability and Computing*, 8:131–159, 1999.

[12] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64, 2001.

[13] K. B. Singer. *Random Intersection Graphs.* PhD thesis, John Hopkins University, Baltimore, Maryland, 1995.

[14] D. Stark. The vertex degree distribution of random intersection graphs. *Random Structures and Algorithms*, 24(3):249–258, May 2004.

[15] R. E. Tarjan and M. Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM Journal on Computing*, 13(3):566–579, 1984.

[16] M. Thimm, A. Goede, S. Hougardy, and R. Preissner. Comparison of 2D similarity and 3D superposition. application to searching a conformational drug database. *Journal of Chemical Information and Computer Sciences*, 44:1816–1822, 2004.

[17] M. Ueckerdt. Färben von zufälligen Schnittgraphen. Master's thesis, Humboldt-Universität zu Berlin, 2005.